

Group Sparse Recovery via the $\ell^0(\ell^2)$ Penalty: Theory and Algorithm

Yuling Jiao, Bangti Jin, Xiliang Lu

Abstract—In this work we propose and analyze a novel approach for group sparse recovery. It is based on regularized least squares with an $\ell^0(\ell^2)$ penalty, which penalizes the number of nonzero groups. One distinct feature of the approach is that it has the built-in decorrelation mechanism within each group, and thus can handle challenging strong inner-group correlation. We provide a complete analysis of the regularized model, e.g., existence of a global minimizer, invariance property, support recovery, and properties of block coordinatewise minimizers. Further, the regularized problem admits an efficient primal dual active set algorithm with a provable finite-step global convergence. At each iteration, it involves solving a least-squares problem on the active set only, and exhibits a fast local convergence, which makes the method extremely efficient for recovering group sparse signals. Extensive numerical experiments are presented to illustrate salient features of the model and the efficiency and accuracy of the algorithm. A comparative study indicates its competitiveness with existing approaches.

Index Terms—group sparsity, block sparsity, blockwise mutual incoherence, global minimizer, block coordinatewise minimizer, primal dual active set algorithm, $\ell^0(\ell^2)$ penalty

I. INTRODUCTION

SPARSE recovery has received much attention in many areas, e.g., signal processing, statistics, and machine learning recently. The key assumption is that the data $y \in \mathbb{R}^n$ is generated by a linear combination of a few atoms of a given dictionary $\Psi \in \mathbb{R}^{n \times p}$, $p \gg n$, where each column represents an atom. In the presence of noise $\eta \in \mathbb{R}^n$ (with a noise level $\epsilon = \|\eta\|$), it is formulated as

$$y = \Psi x^\dagger + \eta, \quad (1)$$

where the vector $x^\dagger \in \mathbb{R}^p$ denotes the signal to be recovered.

The most natural formulation of the problem of finding the sparsest solution is the following ℓ^0 optimization

$$\min_{x \in \mathbb{R}^p} \frac{1}{2} \|\Psi x - y\|^2 + \lambda \|x\|_{\ell^0}, \quad (2)$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector, $\|\cdot\|_{\ell^0}$ denotes the number of nonzero entries, and $\lambda > 0$ is a regularization parameter. Due to discontinuity of the ℓ^0 penalty, it is challenging to find a global minimizer of problem (2). In practice, lasso / basis pursuit [1], [2], which replaces the ℓ^0 penalty with its convex relaxation, the ℓ^1 penalty, has been

very popular. Many deep results on the equivalence between the ℓ^0 and ℓ^1 problems and error estimates have been obtained [3], [4], based on the concepts mutual coherence (MC) and restricted isometry property (RIP).

A. Group sparse recovery

In practice, in addition to sparsity, signals may exhibit additional structure, e.g., nonzero coefficients occur in clusters/groups, which are commonly known as block- / group-sparsity. In electroencephalogram (EEG), each group encodes the information about the direction and strength of the dipoles of each discrete voxel representing the dipole approximation [5]. Other applications include multi-task learning [6], wavelet image analysis [7], [8], gene analysis [9], [10] and multichannel image analysis [11], [12], to name a few. The multiple measurement vector problem is also one special case [13]. In these applications, the focus is to recover all contributing groups, instead of one entry from each group. The group structure is an important piece of *a priori* knowledge about the problem, and should be properly accounted for in the recovery method in order to improve interpretability and accuracy of the recovered signal.

There have been many important developments of group sparse recovery. One popular approach is group lasso, extending lasso using an $\ell^1(\ell^2)$ -penalty [14]–[17]. A number of theoretical studies have shown many desirable properties of group lasso, and its advantages over lasso for recovering group sparse signals [18]–[23] under suitable MC or RIP type conditions. To remedy the drawbacks of group lasso, e.g., biasedness and lack of the oracle property [24], [25], nonconvex penalties have been extended to the group case, e.g., bridge, smoothly clipped absolute deviation (SCAD), and minmax concavity penalty (MCP) [17], [26], [27]. A number of efficient algorithms [16], [28]–[34] have been proposed for convex and nonconvex group sparse recovery models. Like in the sparse case, several group greedy methods have also been developed and analyzed in depth [20], [35], [36].

However, in these interesting works, the submatrices of Ψ are assumed to be well conditioned in order to get estimation errors. While this assumption is reasonable in some applications, it excludes the practically important case of strong correlation within groups. For example, in microarray gene analysis, it was observed that genes in the same pathway produce highly correlated values [37]; in genome-wide association studies, SNPs are highly correlated or even linearly dependent within segments of the DNA sequence [38]; in functional neuroimaging, identifying the brain regions

School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, 430063, P.R. China. (yulingjiaomath@whu.edu.cn)

Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK. (bangti.jin@gmail.com, b.jin@ucl.ac.uk)

Corresponding author. School of Mathematics and Statistics and Hubei Key Laboratory of Computational Science, Wuhan University, Wuhan 430072, P.R. China. (xllv.math@whu.edu.cn)

involved in the cognitive processing of an external stimuli is formalized as identifying the non-zero coefficients of a linear model predicting the external stimuli from the neuroimaging data, where strong correlation occurs between neighboring voxels [39]; just to name a few.

In the presence of strong inner-group correlation, an inadvertent application of standard sparse recovery techniques is unsuitable. Numerically, one often can only recover one predictor within each contributing group, which is undesirable when seeking the whole group [40]. Theoretically, the correlation leads bad RIP or MC conditions, and thus many sparse recovery techniques may perform poorly.

B. The $\ell^0(\ell^2)$ approach and our contributions

In this work, we shall develop and analyze a nonconvex model and algorithm for recovering group-sparse signals with potentially strong inner-group correlation. Our approach is based on the following $\ell^0(\ell^2)$ optimization

$$\min_{x \in \mathbb{R}^p} \{J_\lambda(x) = \frac{1}{2} \|\Psi x - y\|^2 + \lambda \|x\|_{\ell^0(\ell^2)}\}, \quad (3)$$

where the $\ell^0(\ell^2)$ penalty $\|\cdot\|_{\ell^0(\ell^2)}$ (with respect to a given partition $\{G_i\}_{i=1}^N$) is defined below in (6), and the regularization parameter $\lambda > 0$ controls the group sparsity level of the solution. The $\ell^0(\ell^2)$ penalty is to penalize the number of nonzero groups. To the best of our knowledge, this model has not been systematically studied in the literature, even though the $\ell^0(\ell^2)$ penalty was used in several prior works; see Section I-C below. We shall provide both theoretical analysis and efficient solver for the model.

The model (3) has several distinct features. The regularized solution is invariant under full rank column transformation, and does not depend on the specific parametrization within the groups. Thus, it allows strong inner-group correlation and merits a built-in decorrelation effect, and admits theoretical results under very weak conditions. Further, both global minimizer and block coordinatewise minimizer have desirable properties, e.g., support recovery and oracle property.

The main contributions of this work are three-folded. First, we establish fundamental properties of the model (3), e.g., existence of a global minimizer, local optimality, necessary optimality condition, and transformation invariance, which theoretically substantiates (3). For example, the invariance implies that it can be equivalently transformed into a problem with orthonormal columns within each group, and thus it is independent of the conditioning of inner-group columns, which contrasts sharply with most existing group sparse recovery models. Second, we develop an efficient algorithm for solving the model (3), which is of primal dual active set (PDAS) type. It is based on a careful analysis of the necessary optimality system, and represents a nontrivial extension of the PDAS algorithm for the ℓ^1 and ℓ^0 penalties [41], [42]. It is very efficient when coupled with a continuation strategy, due to its Newton nature [41]. Numerically, each inner iteration involves only solving a least-squares problem on the active set. The whole algorithm converges globally in finite steps to the oracle solution. Third, we present extensive numerical experiments to illustrate the features of our approach, and

to show its competitiveness with start-of-art group sparse recovery methods, including group lasso and greedy methods.

C. Connections with existing works and organization

The proposed model (3) is closely related to the following constrained nonconvex optimization

$$\min \|x\|_{\ell^0(\ell^q)} \quad \text{subject to } y = \Psi x, \quad (P_q)$$

in the absence of noise η . This model was studied in [20], [36], [43], [44]. In the case of $q = 2$, Eldar and Mishali [43] discussed unique group sparse recovery, and Eldar et al [20] developed an orthogonal matching pursuit algorithm for recovering group sparse signals and established recovery condition in terms of block coherence. See also [36] for related results for subspace signal separation. Elhamifar and Vidal [44] derived the necessary and sufficient conditions for the equivalence of problem (P_q) with a convex $\ell^1(\ell^q)$ relaxation, and gave sufficient conditions using the concept cumulative subspace coherence. Further, under even weaker conditions, they extended these results to the Ψ -weighted formulation

$$\min \sum_{i=1}^N \|\Psi_{G_i} x_{G_i}\|_{\ell^q}^0 \quad \text{subject to } y = \Psi x, \quad (P'_q)$$

which is especially suitable for redundant dictionaries. The models (P_q) and (P'_q) are equivalent, if the columns within each group are of full column rank. Our approach (3) can be viewed as a natural extension of (P_q) with $q = 2$ to the case of noisy data using a Lagrangian formulation, which, due to the nonconvexity of the $\ell^0(\ell^2)$ penalty, is generally not equivalent to the constrained formulation. In this work, we provide many new insights into analytical properties and algorithm developments for the model (3), which have not been discussed in these prior works. Surprisingly, we shall show that the model (3) has built-in decorrelation effect for redundant dictionaries, similar to the model (P'_q) .

The rest of the paper is organized as follows. In Section II, we describe the problem setting, and derive useful estimates. In Section III, we provide analytical properties, e.g., the existence of a global minimizer, invariance property, and optimality condition. In Section IV, we develop an efficient group primal dual active set with continuation algorithm, and analyze its convergence and computational complexity. Finally, in Section V, several numerical examples are provided to illustrate the mathematical theory and the efficiency of the algorithm. All the technical proofs are given in the appendices.

II. PRELIMINARIES

In this section, we describe the problem setting, and derive useful estimates.

A. Problem setting and notations

Throughout, we assume that the sensing matrix $\Psi \in \mathbb{R}^{n \times p}$ with $n \ll p$ has normalized columns $\|\psi_i\| = 1$ for $i = 1, \dots, p$, and the index set $S = \{1, \dots, p\}$ is divided into N non-overlapping groups $\{G_i\}_{i=1}^N$ such that $1 \leq s_i = |G_i| \leq s$ and $\sum_{i=1}^N |G_i| = p$. For any index set $B \subseteq S$, we denote

by x_B (respectively Ψ_B) the subvector of x (respectively the submatrix of Ψ) which consists of the entries (respectively columns) whose indices are listed in B . All submatrices Ψ_{G_i} , $i = 1, 2, \dots, N$, are assumed to have full column rank. The true signal x^\dagger is assumed to be group sparse with respect to the partition $\{G_i\}_{i=1}^N$, i.e., $x^\dagger = (x_{G_1}^\dagger, \dots, x_{G_N}^\dagger)$, with T nonzero groups. Accordingly, the group index set $\{1, \dots, N\}$ is divided into the active set \mathcal{A}^\dagger and inactive set \mathcal{I}^\dagger by

$$\mathcal{A}^\dagger = \{i : \|x_{G_i}^\dagger\| \neq 0\} \quad \text{and} \quad \mathcal{I}^\dagger = (\mathcal{A}^\dagger)^c. \quad (4)$$

The data vector y in (1), possibly contaminated by noise, can be recast as $y = \Psi x^\dagger + \eta = \sum_{i \in \mathcal{A}^\dagger} \Psi_{G_i} x_{G_i}^\dagger + \eta$. Given the true active set \mathcal{A}^\dagger (as if it were provided by an oracle), we define the oracle solution x^o by the least squares solution on \mathcal{A}^\dagger to (1), i.e.,

$$x^o = \underset{\text{supp}(x) \subseteq \bigcup_{i \in \mathcal{A}^\dagger} G_i}{\text{argmin}} \|\Psi x - y\|^2. \quad (5)$$

The oracle solution x^o is uniquely defined provided that $\Psi_{\bigcup_{i \in \mathcal{A}^\dagger} G_i}$ has full column rank. It is the best approximation for problem (1), and will be used as the benchmark.

For any vector $x \in \mathbb{R}^p$, we define an $\ell^r(\ell^q)$ -penalty (with respect to the partition $\{G_i\}_{i=1}^N$) for $r \geq 0$ and $q > 0$ by

$$\|x\|_{\ell^r(\ell^q)} = \begin{cases} (\sum_{i=1}^N \|x_{G_i}\|_{\ell^q}^r)^{1/r}, & r > 0, \\ \#\{i : \|x_{G_i}\|_{\ell^q} \neq 0\}, & r = 0, \\ \max_i \{\|x_{G_i}\|_{\ell^q}\}, & r = \infty. \end{cases} \quad (6)$$

When $r = q > 0$, the $\ell^r(\ell^q)$ penalty reduces to the usual ℓ^r penalty. The choice $r = 0$ (or $r = \infty$) and $q = 2$ is frequently used below. Further, we shall abuse the notation $\|\cdot\|_{\ell^r(\ell^q)}$ for any vector that is only defined on some subgroups (equivalently zero extension).

For any $r, q \geq 1$, the $\ell^r(\ell^q)$ penalty defines a proper norm, and was studied in [45]. For any $r, q > 0$, the $\ell^r(\ell^q)$ penalty is continuous. The $\ell^0(\ell^2)$ penalty, which is of major interest in this work, is discontinuous, but still lower semi-continuous.

Proposition 1: The $\ell^0(\ell^2)$ penalty is lower semicontinuous.

Proof: Let $\{x^n\} \subset \mathbb{R}^p$ be a convergent sequence to some $x^* \in \mathbb{R}^p$. By the continuity of the ℓ^2 norm, $\|x_{G_i}^n\|$ converges to $\|x_{G_i}^*\|$, for $i = 1, \dots, N$. Now the assertion follows from $\|x_{G_i}^*\|_{\ell^0} \leq \liminf \|x_{G_i}^n\|_{\ell^0}$ [46, Lemma 2.2]. ■

Now we derive the hard-thresholding operator $x^* \in H_\lambda(g)$ for one single group for an s -dimensional vector $g \in \mathbb{R}^s$ as

$$x^* \in \arg \min_{x \in \mathbb{R}^s} \frac{1}{2} \|x - g\|^2 + \lambda \|x\|_{\ell^0(\ell^2)},$$

where the $\|\cdot\|_{\ell^0(\ell^2)}$ penalty is given by $\|x\|_{\ell^0(\ell^2)} = 1$ if $x \neq 0$, and $\|x\|_{\ell^0(\ell^2)} = 0$ otherwise. Then it can be verified directly

$$x^* = \begin{cases} g, & \text{if } \|g\| > \sqrt{2\lambda}, \\ 0, & \text{if } \|g\| < \sqrt{2\lambda}, \\ 0 \text{ or } g, & \text{if } \|g\| = \sqrt{2\lambda}. \end{cases}$$

For a vector $x \in \mathbb{R}^p$, the hard thresholding operator H_λ (with respect to the partition $\{G_i\}_{i=1}^N$) is defined groupwise. For $s = 1$, it recovers the usual hard thresholding operator, and hence it is called a group hard thresholding operator.

B. Blockwise mutual coherence

We shall analyze the model (3) using the concept *blockwise mutual coherence* (BMC). We first introduce some notation:

$$\bar{\Psi}_{G_i} = (\Psi_{G_i}^t \Psi_{G_i})^{\frac{1}{2}} \quad \text{and} \quad D_{i,j} = \bar{\Psi}_{G_i}^{-1} \Psi_{G_i}^t \Psi_{G_j} \bar{\Psi}_{G_j}^{-1}. \quad (7)$$

Since Ψ_{G_i} has full column rank, $\bar{\Psi}_{G_i}$ is symmetric positive definite and invertible.

The main tool in our analysis is the BMC μ of the matrix Ψ with respect to the partition $\{G_i\}_{i=1}^N$, which is defined by

$$\mu = \max_{i \neq j} \mu_{i,j}, \quad \text{where } \mu_{i,j} = \sup_{\substack{u \in \mathcal{N}_i \setminus \{0\} \\ v \in \mathcal{N}_j \setminus \{0\}}} \frac{\langle u, v \rangle}{\|u\| \|v\|}, \quad (8)$$

where \mathcal{N}_i is the subspace spanned by the columns of Ψ_{G_i} , i.e., $\mathcal{N}_i = \text{span}\{\psi_l, l \in G_i\} \subseteq \mathbb{R}^n$. The quantity $\mu_{i,j}$ is the cosine of the minimum angle between two subspaces \mathcal{N}_i and \mathcal{N}_j . Thus the BMC μ generalizes the concept mutual coherence (MC) ν , which is defined by $\nu = \max_{i \neq j} |\langle \psi_i, \psi_j \rangle|$ [47], and is widely used in the analysis of sparse recovery algorithms [42], [48], [49]. The concept BMC was already introduced in [36] for separating subspace signals, and [44] for analyzing convex block sparse recovery. In linear algebra, one often uses principal angles to quantify the angles between two subspaces [50], i.e., given $U, V \subseteq \mathbb{R}^n$, the principal angles θ_l for $l = 1, 2, \dots, \min(\dim U, \dim V)$ are defined recursively by

$$\cos(\theta_l) = \max_{\substack{u \in U, \|u\|=1, u \perp \text{span}\{u_i\}_{i=1}^{l-1} \\ v \in V, \|v\|=1, v \perp \text{span}\{v_j\}_{j=1}^{l-1}}} \langle u, v \rangle.$$

By the definition of principal angles, $\mu_{i,j} = \cos(\theta_1)$ for $(U, V) = (\mathcal{N}_i, \mathcal{N}_j)$; see Lemma 2 below and [50, pp. 603–604] for the proof. Principal angles (and hence BMC) can be computed efficiently by QR and SVD [50], unlike RIP or its variants [51].

Lemma 2: Let $U_i \in \mathbb{R}^{n \times s_i}$ and $V_j \in \mathbb{R}^{n \times s_j}$ be two matrices whose columns are orthonormal basis of \mathcal{N}_i and \mathcal{N}_j , respectively, and $\{\theta_l\}_{l=1}^{\min(s_i, s_j)}$ be the principal angles between \mathcal{N}_i and \mathcal{N}_j . Then, $\mu_{i,j} = \cos(\theta_1) = \sigma_{\max}(U_i^t V_j)$.

The next result shows that the BMC μ can be bounded from above by the MC ν ; see Appendix A for the proof. Hence, the BMC is sharper than a direct extension of the MC, since the BMC does not depend on the inner-group correlation.

Proposition 3: Let the MC ν of Ψ satisfy $(s-1)\nu < 1$. Then for the BMC μ of Ψ , there holds $\mu \leq \frac{\nu s}{1-\nu(s-1)}$.

Below we always assume the following condition.

Assumption 2.1: The BMC μ of Ψ satisfies $\mu \in (0, 1/3T)$. We have a few comments on Assumption 2.1.

Remark 2.1: First, if the group sizes do not vary much, then the condition $\mu < 1/3T$ holds if $\nu < 1/C \|x^\dagger\|_{\ell^0}$. The latter condition with $C \in (2, 7)$ is widely used for analyzing lasso [52] and OMP [49], [53]. Hence, the condition in Assumption 2.1 generalizes the classical one. Second, it allows strong inner-group correlations (i.e., ill-conditioning of Ψ_{G_i}), for which the MC ν can be very close to one, and thus it has a built-in mechanism to tackle inner-group correlation. This differs essentially from existing approaches, which rely on certain pre-processing techniques [54], [55].

Remark 2.2: A similar block MC, defined by $\mu_B = \max_{i \neq j} \|\Psi_{G_i}^t \Psi_{G_j}\|/s$, was used for analyzing group greedy algorithms [20], [35] and group lasso [22] (without scaling s). If every submatrix Ψ_{G_i} is column orthonormal, i.e., $\Psi_{G_i}^t \Psi_{G_i} = I$, then μ_B and μ are identical. However, to obtain the error estimates in [20], [35], the MC ν within each group is still needed, which excludes inner-group correlations. The estimates in [22] were obtained under the assumption $\max_i \|\Psi_{G_i}^t \Psi_{G_i} - I\| \leq 1/2$, which again implies that Ψ_{G_i} are well conditioned [22, Theorem 1]. Group restrict eigenvalue conditions [18], [21] and group RIP [23] were adopted for analyzing the group lasso. Under these conditions, strong correlation within groups is also not allowed.

Now we give a few useful estimates. The proofs can be found in Appendix B.

Lemma 4: For any i, j , there hold

$$\begin{aligned} \|\bar{\Psi}_{G_i}^{-1} \Psi_{G_i}^t y\| &\leq \|y\|, \quad \|\Psi_{G_i} \bar{\Psi}_{G_i}^{-1} x_{G_i}\| = \|x_{G_i}\|, \\ \|D_{i,j} x_{G_j}\| &\begin{cases} \leq \mu \|x_{G_j}\| & i \neq j, \\ = \|x_{G_j}\| & i = j. \end{cases} \end{aligned}$$

Lemma 5: For any distinct groups G_{i_1}, \dots, G_{i_M} , $1 \leq M \leq T$, let

$$D = \begin{pmatrix} D_{i_1, i_1} & \cdots & D_{i_1, i_M} \\ \vdots & \ddots & \vdots \\ D_{i_M, i_1} & \cdots & D_{i_M, i_M} \end{pmatrix} \quad \text{and} \quad x = \begin{pmatrix} x_{G_{i_1}} \\ \vdots \\ x_{G_{i_M}} \end{pmatrix}.$$

There holds $\|Dx\|_{\ell^\infty(\ell^2)} \in [(1 - (M - 1)\mu)\|x\|_{\ell^\infty(\ell^2)}, (1 + (M - 1)\mu)\|x\|_{\ell^\infty(\ell^2)}]$.

Lemma 5 directly implies the uniqueness of the oracle solution x^o ; see Appendix C for the proof.

Corollary 6: If Assumption 2.1 holds, then x^o is unique.

III. THEORY OF THE $\ell^0(\ell^2)$ OPTIMIZATION PROBLEM

Now we analyze the model (3), e.g., existence of a global minimizer, invariance property, support recovery, and block coordinatewise minimizers.

A. Existence and property of a global minimizer

First we show the existence of a global minimizer to problem (3); see Appendix D for the proof.

Theorem 7: There exists a global minimizer to problem (3).

It can be verified directly that the $\ell^0(\ell^2)$ penalty is invariant under group full-rank column transformation, i.e., $\|\bar{\Psi}_{G_i} x_{G_i}\|_{\ell^0} = \|x_{G_i}\|_{\ell^0}$ for nonsingular $\bar{\Psi}_{G_i}$, $i = 1, 2, \dots, N$. Thus problem (3) can be equivalently transformed into

$$\frac{1}{2} \left\| \sum_{i=1}^N \Psi_{G_i} \bar{\Psi}_{G_i}^{-1} \bar{x}_{G_i} - y \right\|^2 + \lambda \|\bar{x}\|_{\ell^0(\ell^2)}. \quad (9)$$

with $\bar{x}_{G_i} = \bar{\Psi}_{G_i} x_{G_i}$. This invariance does not hold for other group sparse penalties, e.g., group lasso and group MCP. Further, the BMC μ is invariant under the transformation, since $\text{span}(\{\psi_l : l \in G_i\}) = \text{span}(\{\Psi_{G_i} \bar{\Psi}_{G_i}^{-1}\}_l)$.

Remark 3.1: Most existing approaches do not distinguish inner- and inter-group columns, and thus require incoherence between the columns within each group in the theoretical analysis. For strong inner-group correlation, a clustering step

is often employed to decorrelate Ψ [54], [55]. In contrast, our approach has a built-in decorrelation mechanism: it is independent of the conditioning of the submatrices $\{\Psi_{G_i}\}_{i=1}^N$.

For a properly chosen λ , a global minimizer has nice properties, e.g., exact support recovery for small noise and oracle property; the proof is given in Appendix E.

Theorem 8: Let Assumption 2.1 hold, x be a global minimizer of (3) with an active set \mathcal{A} , and $\bar{x}_{G_i}^\dagger = \bar{\Psi}_{G_i} x_{G_i}^\dagger$.

- (i) Let $\Lambda = |\{i \in \mathcal{A}^\dagger : \|\bar{x}_{G_i}^\dagger\| < 2\sqrt{2\lambda} + 3\epsilon\}|$. If $\lambda > \epsilon^2/2$, then $|\mathcal{A} \setminus \mathcal{A}^\dagger| + |\mathcal{A}^\dagger \setminus \mathcal{A}| \leq 2\Lambda$.
- (ii) If η is small, i.e., $\epsilon < \min_{i \in \mathcal{A}^\dagger} \{\|\bar{x}_{G_i}^\dagger\|\}/5$, then for any $\lambda \in (\epsilon^2/2, (\min_{i \in \mathcal{A}^\dagger} \{\|\bar{x}_{G_i}^\dagger\|\} - 2\epsilon)^2/8)$, the oracle solution x^o is the only global minimizer to J_λ .

B. Necessary optimality condition

Since problem (3) is highly nonconvex, there seems no convenient characterization of a global minimizer that is amenable with numerical treatment. Hence, we resort to the concept of a block coordinatewise minimizer (BCWM) with respect to the group partition $\{G_i\}_{i=1}^N$, which is minimizing along each group coordinate x_{G_i} [56]. Specifically, a BCWM x^* to the functional J_λ satisfies for $i = 1, 2, \dots, N$

$$x_{G_i}^* \in \arg \min_{x_{G_i} \in \mathbb{R}^{s_i}} J_\lambda(x_{G_1}^*, \dots, x_{G_{i-1}}^*, x_{G_i}, x_{G_{i+1}}^*, \dots, x_{G_N}^*).$$

We have the following necessary and sufficient condition for a BCWM x^* ; see Appendix F for the proof. It is also the necessary optimality condition of a global minimizer x^* .

Theorem 9: The necessary and sufficient optimality condition for a BCWM $x^* \in \mathbb{R}^p$ of problem (3) is given by

$$\bar{x}_{G_i}^* \in H_\lambda(\bar{x}_{G_i}^* + \bar{d}_{G_i}^*), \quad i = 1, \dots, N, \quad (10)$$

where $\bar{x}_{G_i}^* = \bar{\Psi}_{G_i} x_{G_i}^*$, and the dual variable d^* is $d^* = \Psi^t(y - \Psi x^*)$ and $\bar{d}_{G_i}^* = \bar{\Psi}_{G_i}^{-1} d_{G_i}^*$.

Remark 3.2: The optimality system is expressed in terms of the transformed variables \bar{x} and \bar{d} only, instead of the primary variables x and d . This has important consequences for the analysis and algorithm of the $\ell^0(\ell^2)$ model: both should be carried out in the transformed domain. Clearly, (10) is also the optimality system of a BCWM \bar{x}^* for problem (9), concurring with the invariance property.

Notation. In the discussions below, given a primal variable x and dual variable d , we will use (\bar{x}, \bar{d}) for the transformed variables, i.e., $\bar{x}_{G_i} = \bar{\Psi}_{G_i} x_{G_i}$ and $\bar{d}_{G_i} = \bar{\Psi}_{G_i}^{-1} d_{G_i}$, $i = 1, \dots, N$.

Using the group hard-thresholding operator H_λ , we deduce

$$\|\bar{x}_{G_i}^* + \bar{d}_{G_i}^*\| < \sqrt{2\lambda} \Rightarrow \bar{x}_{G_i}^* = 0 \quad (\Leftrightarrow x_{G_i}^* = 0),$$

$$\|\bar{x}_{G_i}^* + \bar{d}_{G_i}^*\| > \sqrt{2\lambda} \Rightarrow \bar{d}_{G_i}^* = 0 \quad (\Leftrightarrow d_{G_i}^* = 0).$$

Combining these two relations gives a simple observation

$$\|\bar{x}_{G_i}\| \geq \sqrt{2\lambda} \geq \|\bar{d}_{G_i}\|. \quad (11)$$

Next we discuss interesting properties of a BCWM x^* . First, it is always a local minimizer, i.e., $J_\lambda(x^* + h) \geq J_\lambda(x^*)$ for all small $h \in \mathbb{R}^p$; see Appendix G for the proof.

Theorem 10: A BCWM x^* of the functional J_λ is a local minimizer. Further, with its active set \mathcal{A} , if $\Psi_{\cup_{i \in \mathcal{A}} G_i}$ has full column rank, then it is a strict local minimizer.

Given the active set \mathcal{A} of a BCWM x^* , if $|\mathcal{A}|$ is controlled, then \mathcal{A} provides information about \mathcal{A}^\dagger ; see Theorem 11 below and Appendix H for the proof. In particular, if the noise η is small, with a proper choice of λ , then $\mathcal{A} \subseteq \mathcal{A}^\dagger$.

Theorem 11: Let Assumption 2.1 hold, and x^* be a BCWM to the model (3) with a support \mathcal{A} and $|\mathcal{A}| \leq T$. Then the following statements hold.

- (i) The inclusion $\{i : \|\bar{x}_{G_i}^\dagger\| \geq 2\sqrt{2\lambda} + 3\epsilon\} \subseteq \mathcal{A}$ holds.
- (ii) The inclusion $\mathcal{A} \subseteq \mathcal{A}^\dagger$ holds if ϵ is small:

$$\epsilon \leq t \min_{i \in \mathcal{A}^\dagger} \{\|\bar{x}_{G_i}^\dagger\|\} \text{ for some } 0 \leq t < \frac{1-3\mu T}{2}. \quad (12)$$

- (iii) If the set $\{i \in \mathcal{A}^\dagger : \|\bar{x}_{G_i}^\dagger\| \in [2\sqrt{2\lambda} - 3\epsilon, 2\sqrt{2\lambda} + 3\epsilon]\}$ is empty, then $\mathcal{A} \subseteq \mathcal{A}^\dagger$.

IV. GROUP PRIMAL-DUAL ACTIVE SET ALGORITHM

Now we develop an efficient, accurate and globally convergent group primal dual active set with continuation (GPDASC) algorithm for problem (3). It generalizes the algorithm for the ℓ^1 and ℓ^0 regularized problems [41], [42] to the group case.

A. GPDASC algorithm

The starting point is the necessary and sufficient optimality condition (10) for a BCWM x^* , cf. Theorem 9. The following two observations from (10) form the basis of the derivation. First, given a BCWM x^* (and its dual variable $d^* = \Psi^t(y - \Psi x^*)$), one can determine the active set \mathcal{A}^* by

$$\mathcal{A}^* = \{i : \|\bar{x}_{G_i}^* + \bar{d}_{G_i}^*\| > \sqrt{2\lambda}\}$$

and the inactive set \mathcal{I}^* its complement, provided that the set $\{i : \|\bar{x}_{G_i}^* + \bar{d}_{G_i}^*\| = \sqrt{2\lambda}\}$ is empty. Second, given the active set \mathcal{A}^* , one can determine uniquely the primal and dual variables x^* and d^* by (with $B = \cup_{i \in \mathcal{A}^*} G_i$)

$$\begin{cases} x_{G_i}^* = 0 \quad \forall i \in \mathcal{I}^* & \text{and} \quad \Psi_B^t \Psi_B x_B^* = \Psi_B^t y, \\ d_{G_j}^* = 0 \quad \forall j \in \mathcal{A}^* & \text{and} \quad d_{G_i}^* = \Psi_{G_i}^t (y - \Psi x^*) \quad \forall i \in \mathcal{I}^*. \end{cases}$$

By iterating these two steps alternately, with the current estimates (x, d) and $(\mathcal{A}, \mathcal{I})$ in place of (x^*, d^*) and $(\mathcal{A}^*, \mathcal{I}^*)$, we arrive at an algorithm for problem (3).

The complete procedure is listed in Algorithm 1. Here $K_{max} \in \mathbb{N}$ is the maximum number of inner iterations, λ_0 is the initial guess of λ . The choice $\lambda_0 = \frac{1}{2}\|y\|^2$ ensures that $x^0 = 0$ is the only global minimizer, cf. Proposition 12 below, with a dual variable $d^0 = \Psi^t y$. The scalar $\rho \in (0, 1)$ is the decreasing factor for λ , which essentially determines the length of the continuation path.

The algorithm consists of two loops: an inner loop of solving problem (3) with a fixed λ using a GPDAS algorithm (lines 6–10), and an outer loop of continuation along the parameter λ by gradually decreasing its value.

In the inner loop, it involves a least-squares problem:

$$x^{k+1} = \underset{\text{supp}(x) \subseteq \cup_{i \in \mathcal{A}_k} G_i}{\text{argmin}} \quad \|\Psi x - y\|,$$

which amounts to solving a (normal) linear system of size $|\cup_{i \in \mathcal{A}_k} G_i| \leq |\mathcal{A}_k|s$. Hence, this step is very efficient, if the active set \mathcal{A}_k is small, which is the case for group sparse

Algorithm 1 GPDASC algorithm

- 1: Input: $\Psi \in \mathbb{R}^{n \times p}$, $\{G_i\}_{i=1}^N$, K_{max} , $\lambda_0 = \frac{1}{2}\|y\|^2$, and $\rho \in (0, 1)$.
- 2: Compute $\bar{\Psi}_{G_i} = (\Psi_{G_i}^t \Psi_{G_i})^{1/2}$.
- 3: Set $x(\lambda_0) = 0$, $d(\lambda_0) = \Psi^t y$, $\mathcal{A}(\lambda_0) = \emptyset$.
- 4: **for** $s = 1, 2, \dots$ **do**
- 5: Set $\lambda_s = \rho \lambda_{s-1}$, $x^0 = x(\lambda_{s-1})$, $d^0 = d(\lambda_{s-1})$, $\mathcal{A}_{-1} = \mathcal{A}(\lambda_{s-1})$.
- 6: **for** $k = 0, 1, \dots, K_{max}$ **do**
- 7: Let $\bar{x}_{G_i}^k = \bar{\Psi}_{G_i} x_{G_i}^k$ and $\bar{d}_{G_i}^k = \bar{\Psi}_{G_i}^{-1} d_{G_i}^k$, and define $\mathcal{A}_k = \{i : \|\bar{x}_{G_i}^k + \bar{d}_{G_i}^k\| > \sqrt{2\lambda_s}\}$.
- 8: Check the stopping criterion $\mathcal{A}_k = \mathcal{A}_{k-1}$.
- 9: Update the primal variable x^{k+1} by
$$x^{k+1} = \underset{\text{supp}(x) \subseteq \cup_{i \in \mathcal{A}_k} G_i}{\text{argmin}} \quad \|\Psi x - y\|.$$
- 10: Update the dual variable by $d^{k+1} = \Psi^t(y - \Psi x^{k+1})$.
- 11: **end for**
- 12: Set the output by $x(\lambda_s)$, $d(\lambda_s)$ and $\mathcal{A}(\lambda_s)$.
- 13: Check the stopping criterion
$$\|\Psi x(\lambda_s) - y\| \leq \epsilon. \quad (13)$$

14: **end for**

signals. Further, since the inner iterates are of Newton type [41], the local convergence should be fast. However, in order to fully exploit this nice feature, a good initial guess of the primal and dual variables (x, d) is required. To this end, we apply a continuation strategy along λ . Specifically, given a large λ_0 , we gradually decrease its value by $\lambda_s = \rho \lambda_{s-1}$, for some decreasing factor $\rho \in (0, 1)$, and take the solution $(x(\lambda_{s-1}), d(\lambda_{s-1}))$ to the λ_{s-1} -problem $J_{\lambda_{s-1}}$ to warm start the λ_s -problem J_{λ_s} .

There are two stopping criteria in the algorithm, at steps 8 and 13, respectively. In the inner loop, one may terminate the iteration if the active set \mathcal{A}_k does not change or a maximum number K_{max} of inner iterations is reached. Since the stopping criterion $\mathcal{A}_k = \mathcal{A}_{k-1}$ for convex optimization may never be reached in the nonconvex context [42], it has to be terminated after a maximum number K_{max} of iterations. Our convergence analysis holds for any $K_{max} \in \mathbb{N}$, including $K_{max} = 1$, and we recommend $K_{max} \leq 5$ in practice. The stopping criterion at step 13 is essentially concerned with the proper choice of λ . The choice of λ stays at the very heart of the model (3). Many rules, e.g., discrepancy principle, balancing principle and information criterion, have been developed for variational regularization [57]. In Algorithm 1, we give only the discrepancy principle (13), assuming that a reliable estimate on the noise level ϵ is available. The rationale behind the principle is that the reconstruction accuracy should be comparable with the data accuracy. Note that the use of (13) (and other rules) does not incur any extra computational overheads, since the sequence of solutions $\{x(\lambda_s)\}$ is already generated along the continuation path.

Now we justify the choice of λ_0 : for large λ , 0 is the only

global minimizer to J_λ . The proof is given in Appendix I.

Proposition 12: The following statements hold.

- (i) For any $\lambda > 0$, $x^* = 0$ is a strict local minimizer to J_λ ;
- (ii) For any $\lambda > \lambda_0 := \frac{1}{2}\|y\|^2$, $x^* = 0$ is the only global minimizer of problem (3).

B. Convergence analysis

Now we state the global convergence of Algorithm 1.

Theorem 13: Let Assumption 2.1 and (12) hold. Then for a proper choice of $\rho \in (0, 1)$, and for any $K_{\max} \geq 1$, Algorithm 1 converges to x^o in a finite number of iterations.

We only sketch the main ideas, and defer the lengthy proof to Appendix J. The most crucial ingredient of the proof is to characterize a monotone decreasing property of the “energy” during the iteration by some auxiliary set Γ_s defined by

$$\Gamma_s = \left\{ i : \|\bar{x}_{G_i}^\dagger\| \geq \sqrt{2s} \right\}. \quad (14)$$

The inclusion $\Gamma_{s_1} \subseteq \Gamma_{s_2}$ holds trivially for $s_1 > s_2$. If \mathcal{A}_k is the active set at the k^{th} iteration, the corresponding energy E_k is defined by $E_k = E(\mathcal{A}_k) = \max_{i \in \mathcal{I}_k} \|\bar{x}_{G_i}^\dagger\|$. Then with properly chosen $s_1 > s_2$, there holds $\Gamma_{s_2\lambda} \subseteq \mathcal{A}_k \subseteq \mathcal{A}^\dagger \Rightarrow \Gamma_{s_2\lambda} \subseteq \mathcal{A}_{k+1} \subseteq \mathcal{A}^\dagger$. This relation characterizes the evolution of the active set \mathcal{A}_k , and provides a crucial strict monotonicity of the energy E_k . This observation is sufficient to show the convergence of the algorithm to the oracle solution x^o in a finite number of steps; see Appendix J for details.

Remark 4.1: The convergence in Theorem 13 holds for any $K_{\max} \in \mathbb{N}$, including $K_{\max} = 1$. According to the proof in Appendix J, the smaller are the factor μT and the noise level ϵ , the smaller is the decreasing factor ρ that one can choose and thus Algorithm 1 takes fewer outer iterations to reach convergence on the continuation path. We often taken $\rho = 0.7$.

C. Computational complexity

Now we comment on the computational complexity of Algorithm 1. First, we consider one inner iteration. Steps 7-8 take $O(p)$ flops. At Step 9, explicitly forming the matrix $\Psi_{B_k}^t \Psi_{B_k}$, $B_k = \cup_{i \in \mathcal{A}_k} G_i$, takes $O(n|B_k|^2)$ flops, and the cost of forming $\Psi^t y$ is negligible since it is often precomputed. The Cholesky factorization costs $O(|B_k|^3)$ flops and the back-substitution needs $O(|B_k|^2)$ flops. Hence step 9 takes $O(\max(|B_k|^3, n|B_k|^2))$ flops. At step 10, the matrix-vector product takes $O(np)$ flops. Hence, the overall cost of one inner iteration is $O(\max(|B_k|^3, pn, n|B_k|^2))$. Since the GPDAS is of Newton type, a few iterations suffice convergence, which is numerically confirmed in Section V. So with a good initial guess, for each fixed λ , the overall cost is $O(\max(|B_k|^3, pn, |B_k|^2 n))$. In particular, if the true solution x^\dagger is sufficiently sparse, i.e., $|B_k| \ll \min(n, \sqrt{p})$, the cost of per inner iteration is $O(np)$.

Generally, one can apply the well-know low-rank Cholesky up/down-date formulas [58] to further reduce the cost. Specifically, with $B_k = \cup_{i \in \mathcal{A}_k} G_i$, we down-date by removing the columns in B_{k-1} but not in B_k at the cost of $O(|B_{k-1} \setminus B_k| |B_{k-1}|^2)$ flops, and update by appending the columns in B_k but not in B_{k-1} in $O(|B_k \setminus B_{k-1}| (|B_{k-1}|^2 + n|B_{k-1}|))$

flops. Then the Cholesky factor of $\Psi_{B_k}^t \Psi_{B_k}$ is $O((|B_{k-1} \cup B_k| - |B_{k-1} \cap B_k|) |B_{k-1}| (n + |B_{k-1}|))$. Along the continuation path, $(|B_{k-1} \cup B_k| - |B_{k-1} \cap B_k|)$ is small, as confirmed in Fig. 5 below, and thus the overall cost is often of $O(np)$.

V. NUMERICAL RESULTS AND DISCUSSIONS

Now we present numerical results to illustrate distinct features of the proposed $\ell^0(\ell^2)$ model and the efficiency and accuracy of Algorithm 1. All the numerical experiments were performed on a four-core desktop computer with 3.16 GHz and 8 GB RAM. The MATLAB code (GPDASC) is available at <http://www0.cs.ucl.ac.uk/staff/b.jin/software/gpdasc.zip>.

A. Experimental setup

First we describe the problem setup of the numerical experiments. In all the numerical examples, the group sparse structure of the true signal x^\dagger is encoded in the partition $\{G_i\}_{i=1}^N$, which is of equal group size s , with $p = Ns$, and x^\dagger has $T = |\mathcal{A}^\dagger|$ nonzero groups. The dynamic range (DR) of the signal x^\dagger is defined by

$$\text{DR} = \frac{\max\{|x_i^\dagger| : x_i^\dagger \neq 0\}}{\min\{|x_i^\dagger| : x_i^\dagger \neq 0\}}.$$

We fix the minimum nonzero entry at $\min\{|x_i^\dagger| : x_i^\dagger \neq 0\} = 1$. The sensing matrix Ψ is constructed as follows. First we generate a random Gaussian matrix $\tilde{\Psi} \in \mathbb{R}^{n \times p}$, $n \ll p$, with its entries following an independent identically distributed (i.i.d.) standard Gaussian distribution with a zero mean and unit variance. Then for any $i \in \{1, 2, \dots, N\}$, we introduce correlation within the i th group G_i by: given $\tilde{\Psi}_{G_i} \in \mathbb{R}^{n \times |G_i|}$ by setting $\bar{\psi}_1 = \tilde{\psi}_1$, $\bar{\psi}_{|G_i|} = \tilde{\psi}_{|G_i|}$ and

$$\bar{\psi}_j = \tilde{\psi}_j + \theta(\tilde{\psi}_{j-1} + \tilde{\psi}_{j+1}), \quad j = 2, \dots, |G_i| - 1,$$

where the parameter $\theta \geq 0$ controls the degree of inner-group correlation: The larger is θ , the stronger is the correlation. Finally, we normalize the matrix $\tilde{\Psi}$ to obtain Ψ such that all columns are of unit length. The data y is formed by adding noise η to the exact data $y^\dagger = \Psi x^\dagger$ componentwise, where the entries η_i follow an i.i.d. Gaussian distribution $N(0, \sigma^2)$. Below we shall denote by the tuple $(n, p, N, T, s, \text{DR}, \theta, \sigma)$ the data generation parameters, and the notation $N_1 : d : N_2$ denotes the sequence of numbers starting with N_1 and less than N_2 with a spacing d .

B. Comparison with existing group sparse models

First we compare our $\ell^0(\ell^2)$ model (3) (and Algorithm 1) with three state-of-the-art group sparse recovery models and algorithms, i.e., group lasso model

$$\min_{x \in \mathbb{R}^p} \|x\|_{\ell^1(\ell^2)} \quad \text{subject to} \quad \|\Psi x - y\| \leq \epsilon$$

(solved by the group SPGL1 method [29], available at <http://www.cs.ubc.ca/~mpf/spgl1/>, last accessed on December 23, 2015), group MCP (GMCP) model [17], [26], [27] (solved by a group coordinate descent (GCD) method [34]), and group OMP (GOMP) [20], [35]. We refer to these references for their implementation details. Since the algorithm

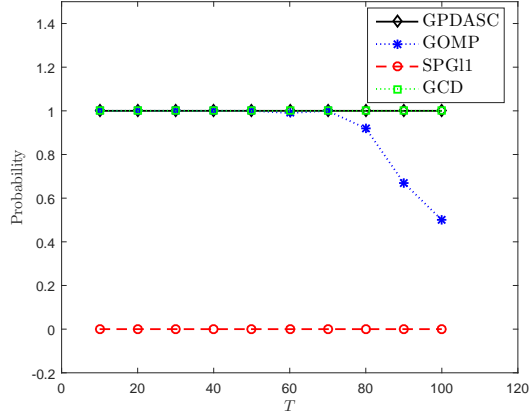
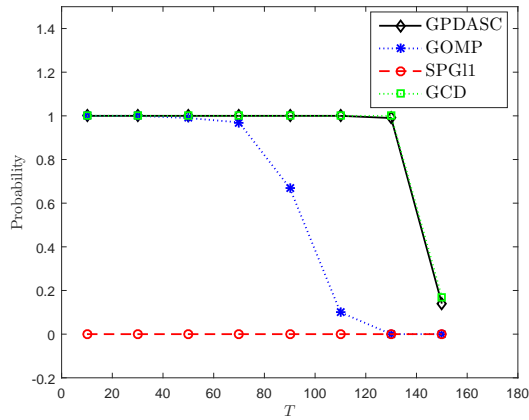
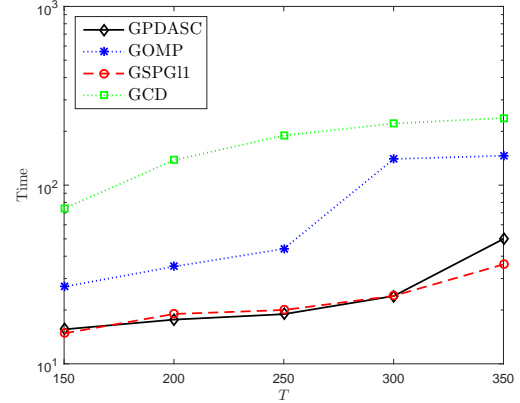
(a) $(800, 2 \times 10^3, 500, 10 : 10 : 100, 4, 10, 0, 10^{-3})$ (b) $(800, 2 \times 10^3, 500, 10 : 10 : 100, 4, 10, 3, 10^{-3})$

Fig. 1: The probability of exact support recovery.

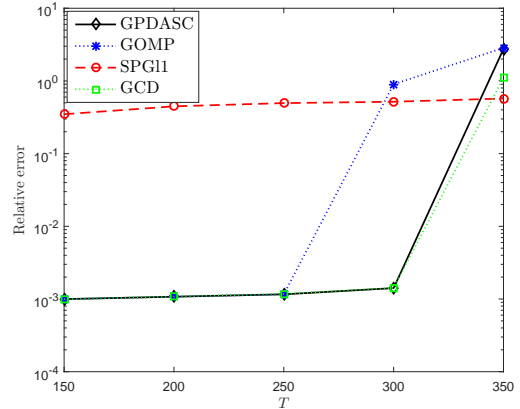
essentially determines the performance of each method, we shall indicate the methods by the respective algorithms, i.e., SPGL1, GCD, GOMP and GPDASC. In the comparison, we examine separately support recovery, and computing time and reconstruction error. All the reported results are the average of 100 independent simulations of the experimental setting.

First, to show exact support recovery, we consider the following two problem settings: $(800, 2 \times 10^3, 500, 10 : 10 : 100, 4, 10, 0, 10^{-3})$ and $(800, 2 \times 10^3, 500, 10 : 10 : 100, 4, 10, 3, 10^{-3})$, for which the condition numbers of the submatrices Ψ_{G_i} are $O(1)$ and $O(10^2)$, respectively, for the case $\theta = 0$ and $\theta = 3$, respectively. Given the group size $s = 4$, the condition number $O(10^2)$ is fairly large, and thus the latter is numerically far more challenging than the former. The numerical results are presented in Fig. 1, where the exact recovery is measured by $\mathcal{A}^* = \mathcal{A}^\dagger$, with \mathcal{A}^\dagger and \mathcal{A}^* being the true and recovered active sets, respectively.

Numerically, it is observed that as the (group) sparsity level T and correlation parameter θ increase, the $\ell^0(\ell^2)$ model and GMCP are the best performers in the test. Theoretically, this is not surprising: the $\ell^0(\ell^2)$ model represents the golden-standard for group sparse recovery, like the ℓ^0 model for the usual sparsity, and GMCP is a close nonconvex proxy to the $\ell^0(\ell^2)$



(a) computing time (in seconds)



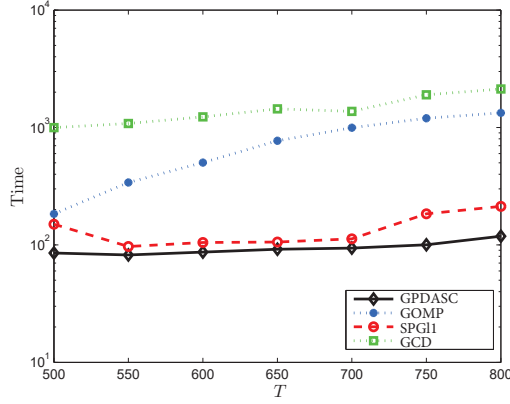
(b) relative error

Fig. 2: Computing time and relative error for GPDASC, GOMP, SPGL1, and GCD for the problem setting $(2 \times 10^3, 1 \times 10^4, 2.5 \times 10^3, 150 : 50 : 350, 4, 100, 1, 10^{-2})$. All computations were performed with the same continuation path.

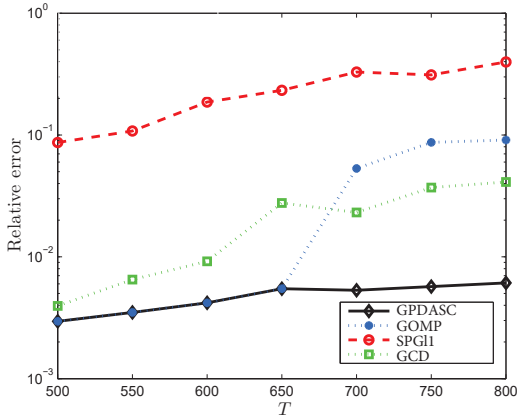
model. Note that GMCP as implemented in [17] is robust with respect to the inner-group correlation, since it performs a preprocessing step to decorrelate Ψ by reorthonormalizing the columns within each group. However, unlike the $\ell^0(\ell^2)$ penalty, this step generally changes the GMCP objective function, due to a lack of transform invariance, and thus may complicate the theoretical analysis of the resulting recovery method. Meanwhile, as a greedy approximation, GOMP does a fairly good job overall: for small θ , it can almost perform as well as the $\ell^0(\ell^2)$ model, but deteriorates greatly for large θ . By its very construction, GOMP from [20] does not take care of the inner-group correlation directly. Surprisingly, group lasso fails most of the time. A closer look at the recovered signals shows that it tends to choose a slightly larger active set than \mathcal{A}^\dagger in the noisy case, and this explains its relatively poor performance in terms of the exact recovery probability, although the relative error is not too large. Intuitively, this concurs with the fact that the convex relaxation often trades the computational efficiency by compromising the reconstruction accuracy.

Next we compare their computing time and reconstruction

error on the following two problem settings: $(2 \times 10^3, 1 \times 10^4, 2.5 \times 10^3, 200 : 25 : 400, 4, 100, 1, 10^{-2})$ and $(5 \times 10^3, 2 \times 10^4, 5 \times 10^3, 500 : 50 : 800, 4, 100, 10, 10^{-3})$, for which the condition number of the submatrices Ψ_{G_i} is of $O(10)$ and $O(10^3)$, respectively. The case $\theta = 10$ involves very strong inner-group correlation, and it is very challenging. The numerical results are presented in Figs. 2 and 3.



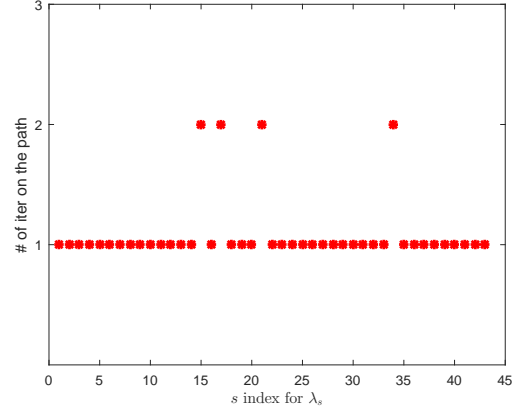
(a) computing time (in second)



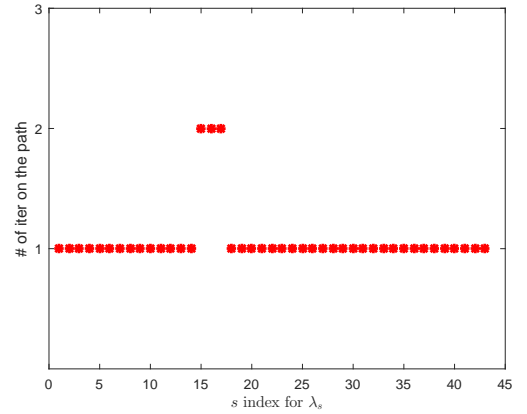
(b) relative error

Fig. 3: Computing time and relative error for GPDASC, GOMP, SPG11, and GCD for the problem setting $(5 \times 10^3, 2 \times 10^4, 5 \times 10^{-3}, 500 : 50 : 800, 4, 100, 10, 10^{-3})$. All computations were performed with the same continuation path.

For $\theta = 1$, the proposed GPDASC for the $\ell^0(\ell^2)$ model is at least three to four times faster than GCD and GOMP, cf. Fig. 2. The efficiency of GPDASC stems from its Newton nature and the continuation strategy, apart from solving least-squares problems only on the active set. We shall examine its convergence more closely below. Group lasso is also computationally attractive, since due to its convexity, it admits an efficient solver SPG11. The coupling with a continuation strategy is beneficial to the efficiency of SPG11 [41]. Meanwhile, the reconstruction errors of the $\ell^0(\ell^2)$ and GMCP are comparable, which is slightly better than GOMP, and they are much accurate than that of group lasso, as observed earlier. In the case of strong inner-group correlation (i.e., $\theta = 10$), the computing time of GPDASC does not change much, but that



(a) $(500, 10^3, 250, 50, 4, 100, 0, 10^{-3})$



(b) $(500, 10^3, 250, 50, 4, 100, 3, 10^{-3})$

Fig. 4: The number of iterations along the continuation path, for each fixed regularization parameter λ_s .

of other algorithms has doubled. Further, the relative error by the $\ell^0(\ell^2)$ model does not deteriorate with the increase of the correlation parameter θ , due to its inherent built-in decorrelation mechanism, cf. Section III, and thus it is far smaller than that by other methods, especially when the group sparsity level T is large. In summary, these experiments show clearly that the proposed $\ell^0(\ell^2)$ model is very competitive in terms of computing time, reconstruction error and exact support recovery.

C. Superlinear local convergence of Algorithm 1

We illustrate the convergence behavior of Algorithm 1 with two problem settings: $(500, 10^3, 250, 50, 4, 100, 0, 10^{-3})$ and $(500, 10^3, 250, 50, 4, 100, 3, 10^{-3})$. To examine the local convergence, we show the number of iterations for each fixed λ_s along the continuation path in Fig. 4. It is observed that the stopping criterion at the inner iteration, i.e., Step 8 of Algorithm 1, is usually reached with one or two iterations, irrespective of the inner-group correlation strength or the regularization parameter λ_s . Hence, Algorithm 1 converges locally superlinearly, like that for the convex ℓ^1 penalty [41], and the continuation strategy can provide a good initial guess for each inner iteration such that the fast local convergence of

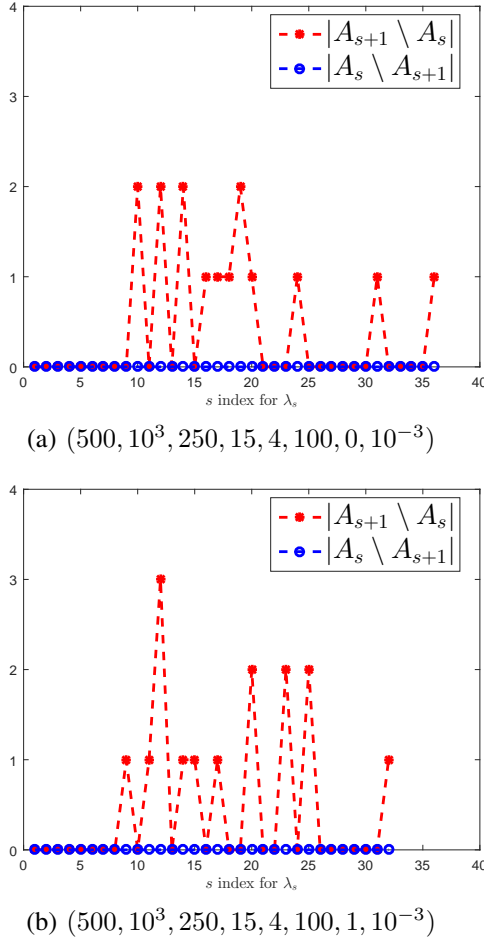


Fig. 5: The variation of the active set size measured by $|A_s \setminus A_{s+1}|$ and $|A_{s+1} \setminus A_s|$ along the continuation path, where A_s denotes the active set at the regularization parameter λ_s .

the GPDAS is fully exploited. This confirms the complexity analysis in Section IV-C. The highly desirable θ -independence convergence is attributed to the built-in de-correlation effect of the $\ell^0(\ell^2)$ model.

To gain further insights, we present in Fig. 5 the variation of the active set along the continuation path using the setting as that of Fig. 4. It is observed that the interesting monotonicity relation $A_s \subset A_{s+1}$ holds along the continuation path. The difference of active sets between two neighboring regularization parameters λ_s is generally small (less than five, and mostly one or two), and thus each GPDAS update is efficient, with a cost comparable with that of one step gradient descent, if using the low-rank Cholesky up/down-date [58], cf. Section IV-C. Further, the empirical observation that each inner iteration often takes only one iteration corroborates the convergence theory in Theorem 13, i.e., the algorithm converges globally even if each inner loop takes one iteration.

Correspondingly, the variation of the relative ℓ^2 error with respect to the oracle solution x^o along the continuation path is given in Fig. 6. For large regularization parameters λ_s , the regularized solution is zero, and thus the relative error is

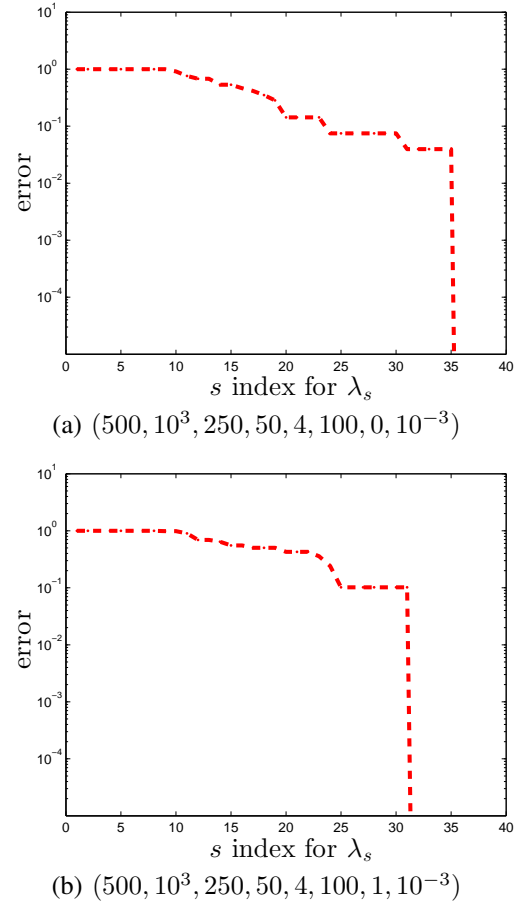


Fig. 6: The relative ℓ^2 error of the iterates along the continuation path, for each fixed regularization parameter λ_s , with respect to the oracle solution x^o .

unit. Then the error first increases slightly, before it starts to decrease monotonically. Upon convergence (i.e., the discrepancy principle is satisfied), the iterate converges to the oracle solution x^o , as indicated by the extremely small error. It is noteworthy that the convergence behavior is almost identical for both the uncorrelated and correlated sensing matrices, further confirming the advantage of the $\ell^0(\ell^2)$ approach.

D. Multichannel image reconstruction

In the last set of experiments, we consider recovering 2D images from compressive and noisy measurement.

The first example is taken from [59]. The target signal is a color image with three-channels $I = (I_r; I_g; I_b)$, with $I_c \in \mathbb{R}^{l^2}$, $c \in \{r, g, b\}$. In the computation, we reorder I into one vector such that the pixels at the same position from the three channels are grouped together. The observational data y is generated by $y = \Psi I + \eta$ where Ψ is a random Gaussian matrix (with correlation within each group) and η is Gaussian noise, following the procedure outlined in Section V-A with the following parameters: $n = 1152$, $p = 6912$, $N = 2304$, $T = 152$, $s = 3$, $\theta = 10$, $\sigma = 1e-3$. The condition number within each group is $O(10^2)$.

The numerical results are presented in Fig. 7 and Table I, where the PSNR is defined by

$$\text{PSNR} = 10 \cdot \log \frac{V^2}{MSE},$$

where V and MSE is the maximum absolute value and the mean squared error, respectively, of the reconstruction. It is observed that GPDASC, GOMP and GCD produce visually equally appealing results, and they are much better than that of SPG11. This observation is also confirmed by the PSNR values in Table I: the PSNR of GPDASC is slightly higher than that of GOMP and GCD. The convergence of GPDASC is much faster than GOMP and GCD. The SPG11 is the most efficient one, but greatly compromises the reconstruction quality.

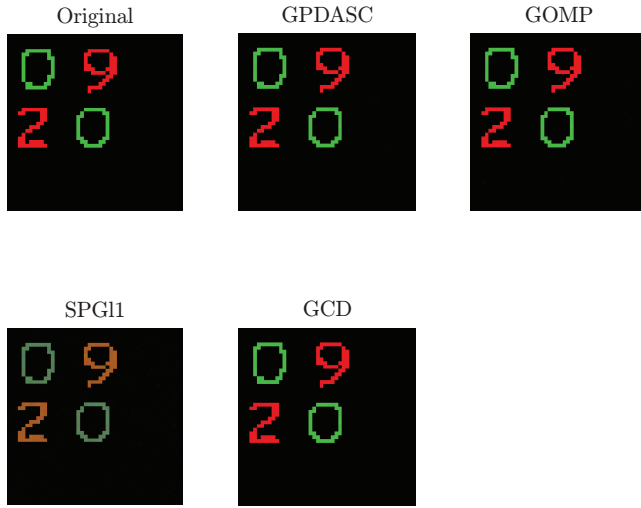


Fig. 7: Reconstruction results of the two-dimensional image.

TABLE I: Numerical results for the two-dimensional image: $n = 1152$, $p = 6912$, $N = 2304$, $T = 152$, $s = 3$, $\theta = 10$, $\sigma = 1e-3$.

algorithm	CPU time (s)	PSNR
GPDASC	5.70	48.2
GOMP	10.9	47.9
SPG11	2.85	22.2
GCD	33.9	48.1

Last, we consider multichannel MRI reconstruction. The sampling matrix Ψ is the composition of a partial FFT with an inverse wavelet transform, with a size 3771×12288 , where we have used 6 levels of Daubechies 1 wavelet. The three channels for each wavelet expansion are organized into one group, and the underlying image $I = (I_r; I_g; I_b)$ has 724 nonzero group coefficients (each of group size 3) under the wavelet transform. Hence, the data is formed as $y = \Psi c + \eta$, where c is the target coefficient with a group sparse structure and η is the Gaussian noise with a noise level $\sigma = 1e-2$. The recovered image I is then obtained by applying the inverse wavelet transform to the estimated coefficient c . The numerical results are presented in Fig. 8 and Table II.

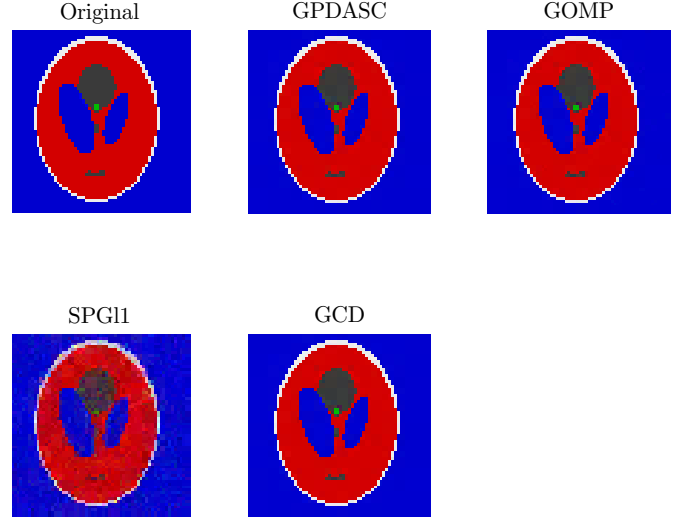


Fig. 8: Reconstructions for the 2D MRI phantom image.

TABLE II: Numerical results for the 2D MRI phantom image: $n = 3771$, $p = 12288$, $N = 4096$, $T = 724$, $s = 3$, $\theta = 0$, $\sigma = 1e-2$.

algorithm	CPU time (s)	PSNR
GPDASC	48.5	38.7
GOMP	203	37.3
SPG11	14.3	20.1
GCD	212	38.2

The observations from the preceding example remain largely valid: the reconstructions by GPDASC, GOMP and GCD are close to each other visually and have comparable PSNR values, and all are much better than that by SPG11. However, GPDASC is a few times faster than that by GOMP and GCD.

VI. CONCLUSIONS

In this work we have proposed and analyzed a novel approach for recovering group sparse signals based on the regularized least-squares problem with an $\ell^0(\ell^2)$ penalty. We provided a complete theoretical analysis on the model, e.g., existence of global minimizers, invariance property, support recovery, and properties of block coordinatewise minimizers. One salient feature of the approach is that it has built-in decorrelation mechanism, and can handle very strong inner-group correlation. Further, these nice properties can be numerically realized efficiently by a primal dual active set solver, for which a finite-step global convergence was also proven. Extensive numerical experiments were presented to illustrate the salient features of the $\ell^0(\ell^2)$ model, and the efficiency and accuracy of the algorithm, and the comparative study with existing approaches show its competitiveness in terms of support recovery, reconstruction errors and computing time.

There are several avenues deserving further study. First, when the column vectors in each group are ill-posed in the sense that they are highly correlated / nearly parallel to each

other, which are characteristic of most inverse problems [60], the proposed $\ell^0(\ell^2)$ model (3) may not be well defined or the involved linear systems in the GPDAS algorithm can be challenging to solve directly. One possible strategy is to apply an extra regularization. This necessitates a refined theoretical study. Second, in practice, the true signal may have extra structure within the group, e.g., smoothness or sparsity. It remains to explore such extra a priori information.

ACKNOWLEDGEMENTS

The authors would like to thank the two referees for their constructive comments. The research of Y. Jiao is partially supported by National Science Foundation of China No. 11501579, B. Jin by EPSRC grant EP/M025160/1, and X. Lu by National Science Foundation of China No. 11471253.

APPENDIX

A. Proof of Proposition 3

Proof: Let $\mathcal{N}_1 = \text{span}\{p_1, \dots, p_{s_1}\}$ and $\mathcal{N}_2 = \text{span}\{q_1, \dots, q_{s_2}\}$ be two subspaces spanned by two distinct groups, where p_i, q_j are column vectors of unit length. By the definition of the MC ν , $|\langle p_i, q_j \rangle| \leq \nu$ for any $i = 1, \dots, s_1$ and $j = 1, \dots, s_2$. For any $u \in \mathcal{N}_1$ and $v \in \mathcal{N}_2$, let $u = \sum_{i=1}^{s_1} c_i p_i$ and $v = \sum_{j=1}^{s_2} d_j q_j$. Then with $c = (c_1, \dots, c_{s_1})$ and $d = (d_1, \dots, d_{s_2})$,

$$\begin{aligned} \|u\|^2 &= \sum_{i,j=1}^{s_1} c_i c_j \langle p_i, p_j \rangle \geq \sum_{i=1}^{s_1} c_i^2 - \nu \sum_{i \neq j} |c_i| |c_j| \\ &\geq (1 - (s_1 - 1)\nu) \|c\|^2 \geq (1 - (s - 1)\nu) \|c\|^2, \end{aligned}$$

and similarly $\|v\|^2 \geq (1 - (s - 1)\nu) \|d\|^2$. Hence we have

$$\frac{|\langle u, v \rangle|}{\|u\| \|v\|} \leq \frac{\nu \sum_{i=1}^{s_1} \sum_{j=1}^{s_2} |c_i d_j|}{(1 - \nu(s - 1)) \|c\| \|d\|} \leq \frac{\nu s}{1 - \nu(s - 1)},$$

by the inequality $\sum_{i=1}^{s_1} \sum_{j=1}^{s_2} |c_i d_j| = \sum_{i=1}^{s_1} |c_i| \sum_{j=1}^{s_2} |d_j| \leq \sqrt{s_1 s_2} \|c\| \|d\| \leq s \|c\| \|d\|$. ■

B. Proof of Lemmas 4 and 5

Proof: [of Lemma 4] First, recall that for any matrix A , $A^t A$ and $A A^t$ have the same nonzero eigenvalues. Upon letting $A = \bar{\Psi}_{G_i}^{-1} \Psi_{G_i}^t$, we have $A A^t = I$, and

$$\|\Psi_{G_i} \bar{\Psi}_{G_i}^{-1} x_{G_i}\|^2 = x_{G_i}^t A A^t x_{G_i} = \|x_{G_i}\|^2,$$

and likewise

$$\|\bar{\Psi}_{G_i}^{-1} \Psi_{G_i}^t y\|^2 = y^t A^t A y \leq \lambda_{\max}(A^t A) \|y\|^2 = \|y\|^2,$$

giving the first two estimates. If $i = j$, $D_{i,j}$ is an identity matrix, and thus $\|D_{i,j} x_{G_j}\| = \|x_{G_j}\|$. For $i \neq j$, $U_i = (\bar{\Psi}_{G_i}^{-1} \Psi_{G_i}^t)^t \in \mathbb{R}^{n \times |s_i|}$, $V_j = (\bar{\Psi}_{G_j}^{-1} \Psi_{G_j}^t)^t \in \mathbb{R}^{n \times |s_j|}$, then

$$D_{i,j} = U_i^t V_j, \quad U_i^t U_i = I, \quad V_j^t V_j = I.$$

Thus by Lemma 2, there holds

$$\|D_{i,j} x_{G_j}\| = \|U_i^t V_j x_{G_j}\| \leq \|U_i^t V_j\| \|x_{G_j}\| \leq \mu \|x_{G_j}\|,$$

showing the last inequality. ■

Proof: [of Lemma 5] Since $D_{i,i} = I$, we have

$$y = Dx = \begin{pmatrix} x_{G_{i_1}} + \sum_{j \neq i_1} D_{i_1, i_j} x_{G_{i_j}} \\ \vdots \\ x_{G_{i_M}} + \sum_{j \neq i_M} D_{i_M, i_j} x_{G_{i_j}} \end{pmatrix} = \begin{pmatrix} y_{G_{i_1}} \\ \vdots \\ y_{G_{i_M}} \end{pmatrix}.$$

By Lemma 4, $\|D_{k, i_j} x_{G_{i_j}}\| \leq \mu \|x_{G_{i_j}}\|$ for any $k \neq i_j$. Let k^* be the index such that $\|y\|_{\ell^\infty(\ell^2)} = \|y_{G_{k^*}}\|$. Then

$$\begin{aligned} \|y\|_{\ell^\infty(\ell^2)} &= \|y_{G_{k^*}}\| \leq \|x_{G_{k^*}}\| + \sum_{i_j \neq k^*} \|D_{k^*, i_j} x_{G_{i_j}}\| \\ &\leq \|x_{G_{k^*}}\| + \mu \sum_{i_j \neq k^*} \|x_{G_{i_j}}\| \leq (1 + (M - 1)\mu) \|x\|_{\ell^\infty(\ell^2)}. \end{aligned}$$

To show the other inequality, let j^* be the index such that $\|x\|_{\ell^\infty(\ell^2)} = \|x_{G_{j^*}}\|$. Then by Lemma 4, we deduce

$$\begin{aligned} \|y\|_{\ell^\infty(\ell^2)} &\geq \|y_{G_{j^*}}\| \geq \|x_{G_{j^*}}\| - \sum_{i_j \neq j^*} \|D_{j^*, i_j} x_{G_{i_j}}\| \\ &\geq \|x_{G_{j^*}}\| - \mu \sum_{i_j \neq j^*} \|x_{G_{i_j}}\| \geq (1 - (M - 1)\mu) \|x\|_{\ell^\infty(\ell^2)}. \end{aligned}$$

This completes the proof of the lemma. ■

C. Proof of Corollary 6

Proof: Since Ψ_{G_i} has full column rank, problem (5) is equivalent to $\bar{x}^o|_{\cup_{i \in \mathcal{A}^\dagger} G_i} = \argmin \|\sum_{i \in \mathcal{A}^\dagger} \Psi_{G_i} \bar{\Psi}_{G_i}^{-1} \bar{x}_{G_i} - y\|^2$, $\bar{x}^o|_{\cup_{i \in \mathcal{I}^\dagger} G_i} = 0$, where $\bar{x}_{G_i} = \bar{\Psi}_{G_i} x_{G_i}$. The normal matrix involved in the least-squares problem on $\cup_{i \in \mathcal{A}^\dagger} G_i$ is exactly the matrix D in Lemma 5, with $\{i_1, \dots, i_M\} = \mathcal{A}^\dagger$. Then the uniqueness of x^o follows from Lemma 5. ■

D. Proof of Theorem 7

Proof: Let $\mathfrak{S} = \{B : B = \cup_{i \in \mathcal{I}} G_i, \mathcal{I} \subseteq \{1, 2, \dots, N\}\}$. Then the set \mathfrak{S} is finite. For any nonempty $B \in \mathfrak{S}$, the problem $\min_{\text{supp}(x) \subseteq B} \|\Psi x - y\|$ has a minimizer $x^*(B)$. Let $T_B^* = \frac{1}{2} \|\Psi x^*(B) - y\|^2 + \lambda \|x^*(B)\|_{\ell^0(\ell^2)}$, and for $B = \emptyset$, let $T_B^* = \frac{1}{2} \|y\|^2$ and $x^*(B) = 0$. Then we denote $T^* = \min_{B \in \mathfrak{S}} T_B^*$, with the minimizing set B^* , and $x^* = x^*(B^*)$. We claim that $J_\lambda(x^*) \leq J_\lambda(x)$ for all $x \in \mathbb{R}^p$. Given any $x \in \mathbb{R}^p$, let $B \in \mathfrak{S}$ be the smallest superset of $\text{supp}(x)$. Then $\|x^*(B)\|_{\ell^0(\ell^2)} \leq \|x\|_{\ell^0(\ell^2)}$, and further by construction $\|\Psi x^*(B) - y\| \leq \|\Psi x - y\|$ and hence $J_\lambda(x) \geq J_\lambda(x^*(B)) \geq J_\lambda(x^*)$. ■

E. Proof of Theorem 8

Proof: Since x^* is a global minimizer of J_λ , we have

$$\lambda T + \frac{1}{2} \epsilon^2 = J_\lambda(x^\dagger) \geq J_\lambda(x^*) \geq \lambda |\mathcal{A}|.$$

This and the choice of λ imply $|\mathcal{A}| \leq T$. Since any global minimizer is also a BCWM, by Theorem 11(i) below, we deduce $\{i \in \mathcal{A}^\dagger : \|\bar{x}_{G_i}^\dagger\| \geq 2\sqrt{2\lambda} + 3\epsilon\} \subseteq \mathcal{A}$. This gives part (i). Next, for $\lambda \in (\epsilon^2/2, (\min_{i \in \mathcal{A}^\dagger} \{\|\bar{x}_{G_i}^\dagger\| - 2\epsilon\}^2/8))$, there holds $\mathcal{A}^\dagger \subseteq \mathcal{A}$ and hence $\mathcal{A}^\dagger = \mathcal{A}$. Hence the only global minimizer is the oracle solution x^o . ■

F. Proof of Theorem 9

Proof: A BCWM x^* is equivalent to the following:

$$x_{G_i}^* \in \operatorname{argmin}_{x_{G_i} \in \mathbb{R}^{s_i}} \frac{1}{2} \|\Psi_{G_i} x_{G_i} + \sum_{j \neq i} \Psi_{G_j} x_{G_j}^* - y\|^2 + \lambda \|x_{G_i}\|_{\ell^0(\ell^2)}$$

for $i = 1, \dots, N$, is equivalent to

$$x_{G_i}^* \in \operatorname{argmin}_{x_{G_i} \in \mathbb{R}^{s_i}} \left\{ \frac{1}{2} \|\Psi_{G_i}(x_{G_i} - x_{G_i}^*)\|^2 + \lambda \|x_{G_i}\|_{\ell^0(\ell^2)} - \langle x_{G_i} - x_{G_i}^*, \Psi_{G_i}^t(y - \Psi x^*) \rangle \right\}.$$

Using the matrices $\bar{\Psi}_{G_i} = (\Psi_{G_i}^t \Psi_{G_i})^{1/2}$ and the identities

$$\begin{cases} \|\Psi_{G_i}(x_{G_i} - x_{G_i}^*)\| = \|\bar{\Psi}_{G_i}(x_{G_i} - x_{G_i}^*)\|, \\ \langle x_{G_i} - x_{G_i}^*, \Psi_{G_i}^t(y - \Psi x^*) \rangle = \langle \bar{\Psi}_{G_i}(x_{G_i} - x_{G_i}^*), \bar{\Psi}_{G_i}^{-1} d_{G_i} \rangle, \\ \|x_{G_i}\|_{\ell^0(\ell^2)} = \|\bar{\Psi}_{G_i} x_{G_i}\|_{\ell^0(\ell^2)}, \end{cases}$$

and recalling $\bar{x}_{G_i} = \bar{\Psi}_{G_i} x_{G_i}$, $\bar{x}_{G_i}^* = \bar{\Psi}_{G_i} x_{G_i}^*$, and $\bar{d}_{G_i}^* = \bar{\Psi}_{G_i}^{-1} d_{G_i}^*$ etc., we deduce

$$\bar{x}_{G_i}^* \in \operatorname{argmin}_{\bar{x}_{G_i} \in \mathbb{R}^{s_i}} \frac{1}{2} \|\bar{x}_{G_i} - (\bar{x}_{G_i}^* + \bar{d}_{G_i}^*)\|^2 + \lambda \|\bar{x}_{G_i}\|_{\ell^0(\ell^2)}.$$

Using the hard-thresholding operator H_λ , we obtain (10). ■

G. Proof of Theorem 10

Proof: It suffices to show $J_\lambda(x^* + h) \geq J_\lambda(x^*)$ for all small $h \in \mathbb{R}^p$. Let $B = \cup_{i \in \mathcal{A}} G_i$. Then

$$x_B^* \in \arg \min \frac{1}{2} \|\Psi_B x_B^* - y\|^2. \quad (15)$$

Now consider a small perturbation $h \in \mathbb{R}^p$ to x^* . If $h_{S \setminus B} = 0$, since $\|x^* + h\|_{\ell^0(\ell^2)} = \|x^*\|_{\ell^0(\ell^2)}$ for small h , by (15), the assertion holds. Otherwise, if $h_{S \setminus B} \neq 0$, then

$$J_\lambda(x^* + h) - J_\lambda(x^*) \geq \lambda - |(h, d^*)|, \quad (16)$$

which is positive for small h , since $\|d^*\|_{\ell^\infty(\ell^2)} \leq \sqrt{2\lambda}$, cf. (11). This shows the first assertion. Now if Ψ_B has full column rank, then problem (15) is strictly convex. Hence, for small $h \neq 0$ with $h_{S \setminus B} = 0$, $\|x^* + h\|_{\ell^0(\ell^2)} = \|x^*\|_{\ell^0(\ell^2)}$ and $\|\Psi(x^* + h) - y\|^2 > \|\Psi x^* - y\|^2$. This and (16) show the second assertion. ■

H. Proof of Theorem 11

First, we derive crucial estimates on one-step primal-dual iteration. Here the energy E associated with an active set \mathcal{A} is defined by

$$E(\mathcal{A}) = \max_{j \in \mathcal{A}^+ \setminus \mathcal{A}} \{\|\bar{x}_{G_j}^\dagger\|\}. \quad (17)$$

These estimates bound the errors in \bar{x} on \mathcal{A} by the energy E and the noise level ϵ , and similarly \bar{d} on \mathcal{I} .

Lemma 14: Let Assumption 2.1 hold, and \mathcal{A} be a given index set with $|\mathcal{A}| \leq T$, and $\mathcal{I} = \mathcal{A}^c$. Consider the following one-step primal-dual update (with $B = \cup_{i \in \mathcal{A}} G_i$)

$$x_B = \Psi_B^\dagger y, \quad x_{S \setminus B} = 0, \quad d = \Psi^t(y - \Psi x), \quad (18)$$

where $\Psi_B^\dagger = (\Psi_B^t \Psi_B)^{-1} \Psi_B^t$ is the pseudo-inverse of Ψ_B . Then with $\mathcal{P} = \mathcal{A} \cap \mathcal{A}^+$, $\mathcal{Q} = \mathcal{A}^+ \setminus \mathcal{A}$ and $\mathcal{R} = \mathcal{A} \setminus \mathcal{A}^+$, $E = E(\mathcal{A})$, for the transformed primal variable \bar{x} , there holds

$$\|\bar{x}_{G_i} - \bar{x}_{G_i}^\dagger\| \leq \frac{1}{1 - |\mathcal{A}|\mu} (|\mathcal{Q}|\mu E + \epsilon) \quad \forall i \in \mathcal{P} \cup \mathcal{R}, \quad (19)$$

and for the transformed dual variable \bar{d} , there holds

$$\begin{aligned} \|\bar{d}_{G_i}\| &\leq C_{|\mathcal{A}|, \mu} (\epsilon + \mu|\mathcal{Q}|E) + |\mathcal{Q}|\mu E + \epsilon, i \in \mathcal{I} \cap \mathcal{I}^+, \\ \|\bar{d}_{G_i}\| &\geq \|\bar{x}_{G_i}^\dagger\| - (C_{|\mathcal{A}|, \mu} (\epsilon + \mu|\mathcal{Q}|E) \\ &\quad + (|\mathcal{Q}| - 1)\mu E + \epsilon), i \in \mathcal{I} \cap \mathcal{A}^+. \end{aligned} \quad (20)$$

with $C_{|\mathcal{A}|, \mu} = |\mathcal{A}|\mu/(1 - \mu(|\mathcal{A}| - 1))$.

Proof: First, the least squares update step in (18) can be rewritten as

$$x_B = (\Psi_B^t \Psi_B)^{-1} \Psi_B^t (\Psi_B x_B^\dagger + \sum_{i \in \mathcal{Q}} \Psi_{G_i} x_{G_i}^\dagger + \eta).$$

Hence, there holds

$$x_B - x_B^\dagger = (\Psi_B^t \Psi_B)^{-1} \Psi_B^t (\sum_{i \in \mathcal{Q}} \Psi_{G_i} x_{G_i}^\dagger + \eta). \quad (21)$$

Let $m = |\mathcal{P}| \leq T$, $\ell = |\mathcal{R}|$, then $k = |\mathcal{Q}| = T - m$. Further, we denote the sets \mathcal{P} , \mathcal{Q} and \mathcal{R} by $\mathcal{P} = \{p_1, \dots, p_m\}$, $\mathcal{Q} = \{q_1, \dots, q_k\}$ and $\mathcal{R} = \{r_1, \dots, r_\ell\}$. Then (21) can be recast blockwise, using $D_{i,j} = \bar{\Psi}_{G_i}^{-1} \Psi_{G_i}^t \Psi_{G_j} \bar{\Psi}_{G_j}^{-1}$ etc, cf. (7), as

$$e := \begin{bmatrix} \bar{x}_{G_{p_1}} - \bar{x}_{G_{p_1}}^\dagger \\ \vdots \\ \bar{x}_{G_{p_m}} - \bar{x}_{G_{p_m}}^\dagger \\ \bar{x}_{G_{r_1}} \\ \vdots \\ \bar{x}_{G_{r_\ell}} \end{bmatrix} = \begin{bmatrix} D_{\mathcal{P}, \mathcal{P}} & D_{\mathcal{P}, \mathcal{R}} \\ D_{\mathcal{R}, \mathcal{R}} & D_{\mathcal{R}, \mathcal{R}} \end{bmatrix} \cdot \left\{ \begin{bmatrix} D_{\mathcal{P}, \mathcal{Q}} \\ D_{\mathcal{R}, \mathcal{Q}} \end{bmatrix} \begin{bmatrix} \bar{x}_{G_{q_1}}^\dagger \\ \vdots \\ \bar{x}_{G_{q_k}}^\dagger \end{bmatrix} + \begin{bmatrix} \bar{\Psi}_{G_{p_1}}^{-1} \Psi_{G_{p_1}}^t \\ \vdots \\ \bar{\Psi}_{G_{p_m}}^{-1} \Psi_{G_{p_m}}^t \\ \bar{\Psi}_{G_{r_1}}^{-1} \Psi_{G_{r_1}}^t \\ \vdots \\ \bar{\Psi}_{G_{r_\ell}}^{-1} \Psi_{G_{r_\ell}}^t \end{bmatrix} \eta \right\},$$

where the matrices $D_{\mathcal{P}, \mathcal{R}}$ etc. are defined by

$$D_{\mathcal{P}, \mathcal{R}} = \begin{bmatrix} D_{p_1, r_1} & \cdots & D_{p_1, r_\ell} \\ \vdots & \vdots & \vdots \\ D_{p_m, r_1} & \cdots & D_{p_m, r_\ell} \end{bmatrix}.$$

Next we estimate the two terms in the curly bracket, denoted by I and II below. By Lemma 4, we deduce

$$\|\text{II}\|_{\ell^\infty(\ell^2)} \leq \|\eta\|. \quad (22)$$

For the first term I, we denote its rows by $z_i = \sum_{j=1}^k D_{i, q_j} \bar{x}_{G_{q_j}}^\dagger$, for any $i \in \mathcal{P} \cup \mathcal{R}$. Since $(\mathcal{P} \cup \mathcal{R}) \cap \mathcal{Q} = \emptyset$, we have for $i \in \mathcal{P} \cup \mathcal{R}$

$$\|z_i\| = \|D_{i, q_1} \bar{x}_{G_{q_1}}^\dagger + \cdots + D_{i, q_k} \bar{x}_{G_{q_k}}^\dagger\| \leq k\mu \max_{1 \leq j \leq k} \{\|\bar{x}_{G_{q_j}}^\dagger\|\}.$$

Since the “energy” $E = \max_{1 \leq j \leq k} \{\|\bar{x}_{G_{aj}}^\dagger\|\}$,

$$\|z\|_{\ell^\infty(\ell^2)} \leq |\mathcal{Q}|\mu E. \quad (23)$$

By Lemma 5, (22) and (23) and the triangle inequality,

$$\|e\|_{\ell^\infty(\ell^2)} \leq \frac{1}{1 - \mu(|\mathcal{P}| + |\mathcal{R}| - 1)} (\epsilon + \mu|\mathcal{Q}|E). \quad (24)$$

Notice that $|\mathcal{P}| + |\mathcal{R}| = |\mathcal{A}|$, we show (19). Next we turn to the transformed dual variable \bar{d} . By the definition, $d = \Psi^t(y - \Psi x)$, and thus for any $i \in \mathcal{I}$, we have

$$d_{G_i} = \Psi_{G_i}^t \left(\sum_{j \in \mathcal{P} \cup \mathcal{R}} \Psi_{G_j}(x_{G_j} - x_{G_j}^\dagger) - \sum_{i \in \mathcal{Q}} \Psi_{G_i} x_{G_i}^\dagger - \eta \right),$$

which upon some algebraic manipulations yields

$$\bar{d}_{G_i} = \sum_{j \in \mathcal{P} \cup \mathcal{R}} D_{i,j}(\bar{x}_{G_j} - \bar{x}_{G_j}^\dagger) - \sum_{j \in \mathcal{Q}} D_{i,j} \bar{x}_{G_j}^\dagger - \bar{\Psi}_{G_i}^{-1} \Psi_{G_i}^t \eta.$$

For any $i \in \mathcal{I} \cap \mathcal{I}^\dagger$, by Lemma 4 and (24), we have

$$\begin{aligned} \|\bar{d}_{G_i}\| &\leq \left\| \sum_{j \in \mathcal{P} \cup \mathcal{R}} D_{i,j}(\bar{x}_{G_j} - \bar{x}_{G_j}^\dagger) \right\| \\ &\quad + \left\| \sum_{j \in \mathcal{Q}} D_{i,j} \bar{x}_{G_j}^\dagger \right\| + \|\bar{\Psi}_{G_i}^{-1} \Psi_{G_i}^t \eta\| \\ &\leq \sum_{j \in \mathcal{P} \cup \mathcal{R}} \mu \|\bar{x}_{G_j} - \bar{x}_{G_j}^\dagger\| + \sum_{j \in \mathcal{Q}} \mu \|\bar{x}_{G_j}^\dagger\| + \epsilon \\ &\leq \frac{(|\mathcal{P}| + |\mathcal{R}|)\mu}{1 - \mu(|\mathcal{P}| + |\mathcal{R}| - 1)} (\epsilon + \mu|\mathcal{Q}|E) + |\mathcal{Q}|\mu E + \epsilon. \end{aligned}$$

The estimate for $i \in \mathcal{I} \cap \mathcal{A}^\dagger = \mathcal{Q}$ follows analogously. ■

Now we can present the proof of Theorem 11.

Proof: First we derive two preliminary estimates using the notation \mathcal{P} , \mathcal{Q} and \mathcal{R} from Lemma 14. Since $|\mathcal{A}| \leq T$ and $|\mathcal{Q}| \leq T$, Lemma 14 and the triangle inequality yield

$$\|\bar{x}_{G_i}\| \leq \frac{1}{1 - T\mu} (T\mu E + \epsilon) \quad \forall i \in \mathcal{A} \cap \mathcal{I}^\dagger. \quad (25)$$

Likewise, using the inequality $\frac{|\mathcal{A}|\mu}{1 - \mu(|\mathcal{A}| - 1)} (\epsilon + \mu|\mathcal{Q}|E) + |\mathcal{Q}|\mu E + \epsilon \leq \frac{1}{1 - T\mu} (T\mu E + \epsilon)$, we deduce from Lemma 14

$$\|\bar{d}_{G_i}\| \geq \|\bar{x}_{G_i}^\dagger\| - \frac{1}{1 - T\mu} (T\mu E + \epsilon) \quad \forall i \in \mathcal{I} \cap \mathcal{A}^\dagger. \quad (26)$$

Now we can proceed to the proof of the theorem. For $\mathcal{Q} = \emptyset$, $\mathcal{A} = \mathcal{A}^\dagger$ and assertions (i) and (ii) are trivially true. Otherwise, let $i^* = \{i \in \mathcal{Q} : \|\bar{x}_{G_i}^\dagger\| = \|\bar{x}_{G_i}^\dagger\|_{\ell^\infty(\ell^2)}\}$. Then $E = \|\bar{x}_{G_{i^*}}^\dagger\|$. By (26) and inequality (11) with $i = i^*$, we have

$$\sqrt{2\lambda} \geq \|\bar{d}_{G_{i^*}}\| \geq E - \frac{T\mu}{1 - T\mu} E - \frac{\epsilon}{1 - T\mu}.$$

Consequently, by Assumption 2.1, we deduce

$$E \leq \frac{1 - T\mu}{1 - 2T\mu} \sqrt{2\lambda} + \frac{1}{1 - 2T\mu} \epsilon < 2\sqrt{2\lambda} + 3\epsilon, \quad (27)$$

i.e., assertion (i) holds. Next we show assertion (ii) by contradiction. If $\mathcal{A} \not\subseteq \mathcal{A}^\dagger$, we can choose $j \in \mathcal{A} \setminus \mathcal{A}^\dagger$, and apply (25) and (26), together with (11), to obtain

$$\frac{1}{1 - T\mu} (T\mu E + \epsilon) \geq \|\bar{x}_{G_j}\| \geq \|\bar{d}_{G_{i^*}}\| \geq \frac{1 - 2T\mu}{1 - T\mu} E - \frac{\epsilon}{1 - T\mu},$$

which contradicts (12), thereby showing assertion (ii). Last, we show assertion (iii). Assume that $\mathcal{A} \not\subseteq \mathcal{A}^\dagger$. Then (27) holds. Meanwhile, since $\mathcal{A} \cap \mathcal{I}^\dagger \neq \emptyset$, using (25) and (26) (by choosing \bar{x}_{G_i} by $i \in \mathcal{A} \cap \mathcal{I}^\dagger$ and $\bar{d}_{G_{i^*}}$) and inequality (11), we have

$$E - \frac{1}{1 - \mu T} (T\mu E + \epsilon) \leq \sqrt{2\lambda} \leq \frac{1}{1 - T\mu} (T\mu E + \epsilon).$$

Under Assumption 2.1, simple computation gives $E \geq 2\sqrt{2\lambda} - 3\epsilon$ and $E \leq 2\sqrt{2\lambda} + 3\epsilon$. This contradicts with the assumption in (iii), and thus the inclusion $\mathcal{A} \subseteq \mathcal{A}^\dagger$ follows. ■

I. Proof of Proposition 12

Proof: Recall the identity $J_\lambda(x) = \frac{1}{2} \|\Psi x - y\|^2 + \lambda \|x\|_{\ell^0(\ell^2)} = J_\lambda(0) + R(x)$, with $R(x) = \frac{1}{2} \|\Psi x\|^2 - \langle \Psi x, y \rangle + \lambda \|x\|_{\ell^0(\ell^2)}$. Also for any $x \neq 0$, $\|x\|_{\ell^0(\ell^2)} \geq 1$. Hence, for any $x \in B_r(0) \setminus \{0\}$, where $B_r(0)$ denotes a ball centered at the origin with a radius $r = \lambda/(\|\Psi^t y\| + 1)$, there holds $R(x) \geq -\|x\| \|\Psi^t y\| + \lambda > 0$. This shows the first assertion. For $\lambda > \lambda_0$, for any nonzero x , we have $\|x\|_{\ell^0(\ell^2)} \geq 1$, and thus $J_\lambda(x) = \frac{1}{2} \|\Psi x - y\|^2 + \lambda \|x\|_{\ell^0(\ell^2)} \geq \lambda > \frac{1}{2} \|y\|^2 = J_\lambda(0)$, i.e., $x^* = 0$ is the only global minimizer. ■

J. Proof of Theorem 13

Proof: The lengthy proof is divided into four steps.

Step 1. First we give the proper choice of the decreasing factor ρ . By (12), we have

$$0 < \frac{1 - \mu T}{1 - 2\mu T - t} < \frac{1 - \mu T}{\mu T + t}.$$

Then for any $s_1 \in ((1 - \mu T)/(1 - 2\mu T - t), (1 - \mu T)/(\mu T + t))$, letting $s_2 = \frac{\mu T + t}{1 - \mu T} s_1 + 1$, we deduce $(1 - \mu T)/(1 - 2\mu T - t) < s_2 < s_1 < (1 - \mu T)/(\mu T + t)$. Combining with the monotonicity of the function $f(s_1) = s_2/s_1$ over the interval $((1 - \mu T)/(1 - 2\mu T - t), (1 - \mu T)/(\mu T + t))$, it implies that for any $\rho \in ((2\mu T + 2t)^2/(1 - \mu T)^2, 1)$, we can find such s_1 with $s_2/s_1 = \sqrt{\rho}$. Next we will choose $\rho \in ((2\mu T + 2t)^2/(1 - \mu T)^2, 1)$.

Step 2. Next we show an important monotonicity relation:

$$\Gamma_{s_1^2 \lambda} \subseteq \mathcal{A}_k \subseteq \mathcal{A}^\dagger \Rightarrow \Gamma_{s_2^2 \lambda} \subseteq \mathcal{A}_{k+1} \subseteq \mathcal{A}^\dagger. \quad (28)$$

For short, we denote by $\mathcal{A} = \mathcal{A}_k$, $\mathcal{I} = \mathcal{I}_k$, and $\mathcal{Q} = \mathcal{A}^\dagger \setminus \mathcal{A}$. By the assumption $\mathcal{A} \subseteq \mathcal{A}^\dagger$, we have $\mathcal{R} = \emptyset$ in Lemma 14. Then it follows from (24) in the proof of Lemma 14 that the updates \bar{x}^{k+1} and \bar{d}^{k+1} satisfy

$$\|\bar{x}_{G_i}^{k+1}\| \geq \|\bar{x}_{G_i}^\dagger\| - \frac{1}{1 - \mu T} (T\mu E_k + \epsilon) \quad \forall i \in \mathcal{A} \quad (29)$$

$$\|\bar{d}_{G_i}^{k+1}\| \leq \frac{1}{1 - \mu T} (T\mu E_k + \epsilon) \quad \forall i \in \mathcal{I}^\dagger, \quad (30)$$

$$\|\bar{d}_{G_i}^{k+1}\| \geq \|\bar{x}_{G_i}^\dagger\| - \frac{1}{1 - \mu T} (T\mu E_k + \epsilon) \quad \forall i \in \mathcal{Q} \quad (31)$$

By the assumption $\Gamma_{s_1^2 \lambda} \subseteq \mathcal{A}_k$, we deduce $E_k < s_1 \sqrt{2\lambda}$; and by assumption (12), $\epsilon < t \min_{i \in \mathcal{A}^\dagger} \{\|\bar{x}_{G_i}^\dagger\|\} \leq t E_k \leq t s_1 \sqrt{2\lambda}$. Hence, using (30), we deduce for any $i \in \mathcal{I}^\dagger$

$$\|\bar{d}_{G_i}^{k+1}\| \leq \frac{1}{1 - \mu T} (T\mu + t) E_k \leq \frac{\mu T + t}{1 - \mu T} s_1 \sqrt{2\lambda} < \sqrt{2\lambda},$$

where the last inequality follows from the choice of s_1 . This and the relation (11) imply that $i \in \mathcal{I}_{k+1}$, and thus $\mathcal{A}_{k+1} \subseteq \mathcal{A}^\dagger$. Meanwhile, by (31), for any $i \in \mathcal{I} \cap \Gamma_{s_2\lambda}$, we have

$$\begin{aligned} \|\bar{d}_{G_i}^{k+1}\| &\geq s_2\sqrt{2\lambda} - \frac{1}{1-\mu T}(\mu T + t)s_1\sqrt{2\lambda} \\ &\geq (s_2 - \frac{\mu T + t}{1-\mu T}s_1)\sqrt{2\lambda} > \sqrt{2\lambda}. \end{aligned}$$

which by the relation (11) yields $i \in \mathcal{A}_{k+1}$. It remains to show $\mathcal{A} \cap \Gamma_{s_2\lambda} \subseteq \mathcal{A}_{k+1}$. Clearly, if $\mathcal{A} = \emptyset$, the assertion is true. Otherwise, for any $i \in \mathcal{A} \cap \Gamma_{s_2\lambda}$, by (29), there holds

$$\begin{aligned} \|\bar{x}_{G_i}\| &\geq \|\bar{x}_{G_i}^\dagger\| - \frac{|Q|\mu + t}{1-(T-1)\mu}\|x_Q^\dagger\|_{\ell^\infty(\ell^2)} \\ &> s_2\sqrt{2\lambda} - \frac{(T-1)\mu + t}{1-T\mu}s_1\sqrt{2\lambda} \geq \sqrt{2\lambda}. \end{aligned}$$

Like before, this and (11) also imply $i \in \mathcal{A}_{k+1}$. Hence the inclusion $\Gamma_{s_2\lambda} \subseteq \mathcal{A}_{k+1}$ holds.

Step 3. Now we prove that the oracle solution x^o is achieved along the continuation path, i.e., $\mathcal{A}(\lambda_s) = \mathcal{A}^\dagger$ for some λ_s . For each λ_s -problem J_{λ_s} , we denote by $\mathcal{A}_{s,0}$ and $\mathcal{A}_{s,\diamond}$ the active set for the initial guess and the last inner step (i.e., $\mathcal{A}(\lambda_s)$ in Algorithm 1) of the s th iterate of the outer loop, respectively. Since $s_1 > s_2$, the inclusion $\Gamma_{s_1\lambda_s} \subseteq \Gamma_{s_2\lambda_s}$ holds. Next we claim that the following inclusion by mathematical induction

$$\Gamma_{s_1\lambda_s} \subseteq \mathcal{A}(\lambda_s) \subseteq \mathcal{A}^\dagger$$

holds for the sequence active sets $\mathcal{A}(\lambda_s)$ from Algorithm 1. From (28), for any index s before the stopping criterion at step 13 of Algorithm 1 is reached, there hold

$$\Gamma_{s_1\lambda_s} \subseteq \mathcal{A}_{s,0} \quad \text{and} \quad \Gamma_{s_2\lambda_s} \subseteq \mathcal{A}_{s,\diamond}. \quad (32)$$

Note that for $s = 0$, by the choice of λ_0 , $\Gamma_{s_1\lambda_0} = \Gamma_{s_2\lambda_0} = \emptyset$, and thus (32) holds. Now for $s > 0$, it follows by mathematical induction and the relation $\mathcal{A}_{s,\diamond} = \mathcal{A}_{s+1,0}$. By (32), during the iteration, the active set $\mathcal{A}_{s,\diamond}$ always lies in \mathcal{A}^\dagger . This shows the desired claim. For large s , we have $\Gamma_{s_1\lambda_s} = \mathcal{A}^\dagger$, and hence $\mathcal{A}(\lambda_s) = \mathcal{A}^\dagger$, and accordingly $x(\lambda_s)$ is the oracle solution x^o .

Step 4. Last, at this step we show that if $\mathcal{A}(\lambda_s) \subsetneq \mathcal{A}^\dagger$, then the stopping criterion at step 13 of Algorithm 1 cannot be satisfied. Let $\mathcal{P} = \mathcal{A}(\lambda_s) \subsetneq \mathcal{A}^\dagger$ and $\mathcal{Q} = \mathcal{A}^\dagger \setminus \mathcal{A}$, and denote by $i^* = \arg \max_{i \in \mathcal{Q}} \{\|\bar{x}_{G_i}^\dagger\|\}$ and $E = \|\bar{x}_{G_{i^*}}^\dagger\|$. Then with the notation $\bar{\Psi}_{G_i}$ and $D_{i,j}$ etc. from (7), we deduce

$$\begin{aligned} \|\Psi x - y\|^2 &= \left\| \sum_{i \in \mathcal{P}} \Psi_{G_i}(x_{G_i} - x_{G_i}^\dagger) - \sum_{j \in \mathcal{Q}} \Psi_{G_j}x_{G_j}^\dagger - \eta \right\|^2 \\ &\geq \|\Psi_{G_{i^*}}x_{G_{i^*}}^\dagger\|^2 + 2 \sum_{j \in \mathcal{Q} \setminus \{i^*\}} \langle \Psi_{G_j}x_{G_j}^\dagger, \Psi_{G_{i^*}}x_{G_{i^*}}^\dagger \rangle \\ &\quad - 2 \sum_{i \in \mathcal{P}} \langle \Psi_{G_i}(x_{G_i} - x_{G_i}^\dagger), \Psi_{G_{i^*}}x_{G_{i^*}}^\dagger \rangle + 2 \langle \eta, \Psi_{G_{i^*}}x_{G_{i^*}}^\dagger \rangle. \end{aligned}$$

Now recall the elementary identities $\|\Psi_{G_{i^*}}x_{G_{i^*}}^\dagger\| = \|\bar{x}_{G_{i^*}}\|$ and $\langle \Psi_{G_j}x_{G_j}^\dagger, \Psi_{G_{i^*}}x_{G_{i^*}}^\dagger \rangle = \langle D_{i^*,j}\bar{x}_{G_i}^\dagger, \bar{x}_{G_{i^*}}^\dagger \rangle$ and then ap-

pealing to Lemma 5, we arrive at

$$\begin{aligned} \|\Psi x - y\|^2 &\geq \|\bar{x}_{G_{i^*}}^\dagger\|^2 + 2 \sum_{j \in \mathcal{Q} \setminus \{i^*\}} \langle D_{i^*,j}\bar{x}_{G_j}^\dagger, \bar{x}_{G_{i^*}}^\dagger \rangle \\ &\quad - 2 \sum_{i \in \mathcal{P}} \langle D_{i^*,i}(\bar{x}_{G_i} - \bar{x}_{G_i}^\dagger), \bar{x}_{G_{i^*}}^\dagger \rangle + 2 \langle \eta, \Psi_{G_{i^*}}x_{G_{i^*}}^\dagger \rangle \\ &\geq E^2 - 2\mu(|\mathcal{Q}| - 1)E^2 - 2\mu|\mathcal{P}|E \max_{i \in \mathcal{P}} \|\bar{x}_{G_i} - \bar{x}_{G_i}^\dagger\| - 2\epsilon E. \end{aligned}$$

By repeating the proof of Lemma 14, we deduce

$$\max_{i \in \mathcal{P}} \|\bar{x}_{G_i} - \bar{x}_{G_i}^\dagger\| \leq \frac{\epsilon + \mu TE}{1 - \mu T}.$$

By assumption (12), $\epsilon \leq tE$, it suffices to show

$$E^2 - 2\mu(|\mathcal{Q}| - 1)E^2 - 2\mu(T - |\mathcal{Q}|)\frac{t + \mu T}{1 - \mu T}E^2 - 2tE^2 > t^2E^2, \quad (33)$$

which implies that the stopping criterion (13) at step 13 of Algorithm 1 cannot be satisfied. The left hand side of (33) is a function monotonically decreasing with respect to the length $|\mathcal{Q}|$, and when $|\mathcal{Q}| = T$, we have $1 - \mu(T - 1) - 2t > t > t^2$, which completes the proof. ■

REFERENCES

- [1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.
- [2] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.
- [3] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.
- [4] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, 2006.
- [5] W. Ou, M. S. Hämmäläinen, and P. Golland, "A distributed spatio-temporal EEG/MEG inverse solver," *NeuroImage*, vol. 44, no. 3, pp. 932–946, 2009.
- [6] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Mach. Learn.*, vol. 73, no. 3, pp. 243–272, 2008.
- [7] J. M. Shapiro, "Embedded image coding using zero-trees of wavelet coefficients," *IEEE Trans. Signal Proc.*, vol. 41, no. 12, pp. 3445–3462, 1993.
- [8] A. Antoniadis and J. Fan, "Regularization of wavelet approximations," *J. Amer. Stat. Assoc.*, vol. 96, no. 455, pp. 939–967, 2001.
- [9] Z. He and W. Yu, "Stable feature selection for biomarker discovery," *Comput. Biol. Chem.*, vol. 34, no. 4, pp. 215–225, 2010.
- [10] S. Ma, X. Song, and J. Huang, "Supervised group lasso with applications to microarray data analysis," *BMC Bioinform.*, vol. 8, no. 1, pp. 60, 17 pp., 2007.
- [11] M. Mishali and Y. C. Eldar, "Blind multiband signal reconstruction: Compressed sensing for analog signals," *IEEE Trans. Signal Proc.*, vol. 57, no. 3, pp. 993–1009, 2009.
- [12] —, "From theory to practice: Sub-nyquist sampling of sparse wide-band analog signals," *IEEE J. Sel. Topics Signal Proc.*, vol. 4, no. 2, pp. 375–391, 2010.
- [13] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Trans. Signal Proc.*, vol. 54, no. 12, pp. 4634–4643, 2006.
- [14] S. Bakin, "Adaptive Regression and Model Selection in Data Mining Problems," Ph.D. dissertation, The Australian National University, 1999.
- [15] D. Malioutov, M. Çetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Proc.*, vol. 53, no. 8, pp. 3704–3716, 2005.
- [16] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. R. Stat. Soc. Ser. B*, vol. 68, no. 1, pp. 49–67, 2006.
- [17] J. Huang, P. Breheny, and S. Ma, "A selective review of group selection in high-dimensional models," *Stat. Sci.*, vol. 27, no. 4, pp. 481–499, 2012.
- [18] J. Huang and T. Zhang, "The benefit of group sparsity," *Ann. Stat.*, vol. 38, no. 4, pp. 1978–2004, 2010.

- [19] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inform. Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.
- [20] Y. C. Eldar, P. Kuppinger, and H. Bölcskei, "Block-sparse signals: uncertainty relations and efficient recovery," *IEEE Trans. Signal Proc.*, vol. 58, no. 6, pp. 3042–3054, 2010.
- [21] K. Lounici, M. Pontil, S. van de Geer, and A. B. Tsybakov, "Oracle inequalities and optimal inference under group sparsity," *Ann. Stat.*, vol. 39, no. 4, pp. 2164–2204, 2011.
- [22] W. U. Bajwa, M. F. Duarte, and R. Calderbank, "Conditioning of random block dictionaries with applications to block-sparse recovery and regression," *IEEE Trans. Inform. Theory*, vol. 61, no. 7, pp. 4060–4079, 2015.
- [23] M. Eren Ahsen and M. Vidyasagar, "Error bounds for compressed sensing algorithms with group sparsity: A unified approach," *Appl. Comput. Harmon. Anal.*, p. in press, 2015.
- [24] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Stat. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.
- [25] C.-H. Zhang and T. Zhang, "A general theory of concave regularization for high-dimensional sparse estimation problems," *Stat. Sci.*, vol. 27, no. 4, pp. 576–593, 2012.
- [26] L. Wang, H. Li, and J. Z. Huang, "Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements," *J. Amer. Stat. Assoc.*, vol. 103, no. 484, pp. 1556–1569, 2008.
- [27] J. Huang, S. Ma, H. Xie, and C.-H. Zhang, "A group bridge approach for variable selection," *Biometrika*, vol. 96, no. 4, p. 1024, 2009.
- [28] L. Meier, S. van de Geer, and P. Bühlmann, "The group Lasso for logistic regression," *J. R. Stat. Soc. Ser. B*, vol. 70, no. 1, pp. 53–71, 2008.
- [29] E. Van Den Berg and M. P. Friedlander, "Probing the Pareto frontier for basis pursuit solutions," *SIAM J. Sci. Comput.*, vol. 31, no. 2, pp. 890–912, 2008.
- [30] P. Tseng and S. Yun, "A coordinate gradient descent method for nonsmooth separable minimization," *Math. Program.*, vol. 117, no. 1–2, Ser. B, pp. 387–423, 2009.
- [31] X. Chen, Q. Lin, S. Kim, J. G. Carbonell, and E. P. Xing, "Smoothing proximal gradient method for general structured sparse regression," *Ann. Appl. Stat.*, vol. 6, no. 2, pp. 719–752, 2012.
- [32] Y. She, "An iterative algorithm for fitting nonconvex penalized generalized linear models with group predictors," *Comput. Stat. Data Anal.*, vol. 56, no. 10, pp. 2976–2990, 2012.
- [33] Z. Qin, K. Scheinberg, and D. Goldfarb, "Efficient block-coordinate descent algorithms for the group Lasso," *Math. Program. Comput.*, vol. 5, no. 2, pp. 143–169, 2013.
- [34] P. Breheny and J. Huang, "Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors," *Stat. Comput.*, vol. 25, no. 2, pp. 173–187, 2015.
- [35] Z. Ben-Haim and Y. C. Eldar, "Near-oracle performance of greedy block-sparse estimation techniques from noisy measurements," *IEEE J. Sel. Topics Signal Proc.*, vol. 5, no. 5, pp. 1032–1047, 2011.
- [36] A. Ganesh, Z. Zhou, and Y. Ma, "Separation of a subspace-sparse signal: algorithms and conditions," *ICASSP 2009*, pp. 3141–3144, 2009.
- [37] M. R. Segal, K. D. Dahlquist, and B. R. Conklin, "Regression approaches for microarray data analysis," *J. Comput. Biol.*, vol. 10, no. 6, pp. 961–980, 2004.
- [38] D. J. Balding, "A tutorial on statistical methods for population association studies," *Nature Reviews Genetics*, vol. 7, no. 10, pp. 781–791, 2006.
- [39] S. M. Tom, C. R. Fox, C. Trepel, and R. A. Poldrack, "The neural basis of loss aversion in decision-making under risk," *Science*, vol. 315, no. 5811, pp. 515–518, 2007.
- [40] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005.
- [41] Q. Fan, Y. Jiao, and X. Lu, "A primal dual active set algorithm with continuation for compressed sensing," *IEEE Trans. Signal Process.*, vol. 62, no. 23, pp. 6276–6285, 2014.
- [42] Y. Jiao, B. Jin, and X. Lu, "A primal dual active set with continuation algorithm for the ℓ^0 -regularized optimization problem," *Appl. Comput. Harmon. Anal.*, vol. 39, no. 3, pp. 400–426, 2015.
- [43] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.
- [44] E. Elhamifar and R. Vidal, "Block sparse recovery via convex optimization," *IEEE Trans. Signal Proc.*, vol. 60, no. 8, pp. 4094–4107, 2012.
- [45] M. Kowalski, "Sparse regression using mixed norms," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 303–324, 2009.
- [46] K. Ito and K. Kunisch, "A variational approach to sparsity optimization based on Lagrange multiplier theory," *Inverse Problems*, vol. 30, no. 1, pp. 015001, 23 pp., 2014.
- [47] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inform. Theory*, vol. 47, no. 7, pp. 2845–2862, 2001.
- [48] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.
- [49] T. T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," *IEEE Trans. Inform. Theory*, vol. 57, no. 7, pp. 4680–4688, 2011.
- [50] Å. Björck and G. H. Golub, "Numerical methods for computing angles between linear subspaces," *Math. Comp.*, vol. 27, pp. 579–594, 1973.
- [51] A. S. Bandeira, E. Dobriban, D. G. Mixon, and W. F. Sawin, "Certifying the restricted isometry property is hard," *IEEE Trans. Inform. Theory*, vol. 59, no. 6, pp. 3448–3450, 2013.
- [52] T. Zhang, "Some sharp performance bounds for least squares regression with L_1 regularization," *Ann. Stat.*, vol. 37, no. 5A, pp. 2109–2144, 2009.
- [53] J. A. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.
- [54] P. Bühlmann, P. Rütimann, S. van de Geer, and C.-H. Zhang, "Correlated variables in regression: clustering and sparse estimation," *J. Statist. Plann. Inference*, vol. 143, no. 11, pp. 1835–1858, 2013.
- [55] D. M. Witten, A. Shojaie, and F. Zhang, "The cluster elastic net for high-dimensional regression with unknown variable grouping," *Technometrics*, vol. 56, no. 1, pp. 112–122, 2014.
- [56] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *J. Optim. Theory Appl.*, vol. 109, no. 3, pp. 475–494, 2001.
- [57] K. Ito and B. Jin, *Inverse Problems: Tikhonov Theory and Algorithms*. World Scientific, NJ, 2014.
- [58] P. E. Gill, G. H. Golub, W. Murray, and M. A. Saunders, "Methods for modifying matrix factorizations," *Math. Comp.*, vol. 28, pp. 505–535, 1974.
- [59] J. Huang, X. Huang, and D. Metaxas, "Learning with dynamic group sparsity," in *2009 IEEE 12th ICCV*, 2009, pp. 64–71.
- [60] G. S. Alverti, H. Ammari, B. Jin, J.-K. Seo, and W. Zhang, "The linearized inverse problem in multifrequency electrical impedance tomography," *SIAM J. Imag. Sci.*, vol. 9, no. 4, pp. 1525–1551, 2016.